



University of Pennsylvania  
**ScholarlyCommons**

---

Publicly Accessible Penn Dissertations


---

2021

## Single-Cell Lineage Tracing Of Cancer Metastasis

Kamen Simeonov  
*University of Pennsylvania*

Follow this and additional works at: <https://repository.upenn.edu/edissertations>

 Part of the [Cell Biology Commons](#), and the [Molecular Biology Commons](#)

---

### Recommended Citation

Simeonov, Kamen, "Single-Cell Lineage Tracing Of Cancer Metastasis" (2021). *Publicly Accessible Penn Dissertations*. 5269.

<https://repository.upenn.edu/edissertations/5269>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/edissertations/5269>  
For more information, please contact [repository@pobox.upenn.edu](mailto:repository@pobox.upenn.edu).

---

## Single-Cell Lineage Tracing Of Cancer Metastasis

### Abstract

The underpinnings of cancer metastasis remain poorly understood, in part due to a lack of tools for probing their emergence at high resolution. Here we present macsGESTALT, an inducible CRISPR-Cas9-based lineage recorder with highly efficient single-cell capture of both transcriptional and phylogenetic information. Applying macsGESTALT to a mouse model of metastatic pancreatic cancer, we recover ~380,000 CRISPR target sites and reconstruct dissemination of ~28,000 single cells across multiple metastatic sites. We find cells occupy a continuum of epithelial-to-mesenchymal transition (EMT) states. Metastatic potential peaks in rare, late-hybrid EMT states, which are aggressively selected from a predominately epithelial ancestral pool. The gene signatures of these late-hybrid EMT states are predictive of reduced survival in both human pancreatic and lung cancer patients, highlighting their relevance to clinical disease progression. Finally, we observe evidence for in vivo propagation of S100 family gene expression across clonally distinct metastatic subpopulations.

### Degree Type

Dissertation

### Degree Name

Doctor of Philosophy (PhD)

### Graduate Group

Cell & Molecular Biology

### First Advisor

Christopher Lengner

### Subject Categories

Cell Biology | Molecular Biology

SINGLE-CELL LINEAGE TRACING OF CANCER METASTASIS

Kamen P. Simeonov

A DISSERTATION

in

Cell and Molecular Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

---

Christopher J. Lengner, PhD  
Associate Professor of Biomedical Sciences

Graduate Group Chairperson

---

Daniel S. Kessler, PhD  
Associate Professor of Cell and Developmental Biology

Dissertation Committee

John I. Murray, PhD  
Associate Professor of Genetics

Ben Z. Stanger, MD, PhD  
Hanna Wise Professor in Cancer Research

Arjun Raj, PhD  
Professor of Genetics

Kai Tan, PhD  
Professor of Pediatrics

# SINGLE-CELL LINEAGE TRACING OF CANCER METASTASIS

COPYRIGHT

2021

Kamen P. Simeonov

*For Peter, Petia, and Kitini*

## ACKNOWLEDGMENT

In order to perform research, one must search and then search again – a quote I first heard from my mom, underscoring the stubborn persistence required to be a scientist. The persistence that enabled this thesis work was only possible with the support of many over the years, both personally and professionally.

Thank you to my parents, Peter and Petia, who immigrated our family to the United States from Bulgaria during the 1990s with nothing, and yet provided every opportunity imaginable for me to succeed. Thank you for challenging me to work hard, while providing unconditional love. And I'm not sure how these things work, but it's almost certainly because of both of you – thank you for instilling in me, an endless curiosity of the natural world and a love of science.

To my best friend and girlfriend, China, we started this long program together to the day, and it's looking like we'll end it together too. In case I ever forget my parents' example, I can always look to you as an example of persistence and hard work. Thank you for supporting me tirelessly, for your unrivaled spontaneity, and for somehow enjoying when I summarize entire movie plots to you more than actually watching the movie.

To the unique friends I've found over the years: Danny, Daniel, Josh, Grant, and Meg. Thank you for being both the funniest and most trustworthy people that exist. Thank you especially to Meg for being impossibly generous and for making work indistinguishable from a comedy podcast. Incidentally, Daniel, Meg, and China are all excellent scientists,

that I've had the pleasure of working with on various projects over the years and hope to again and again.

Thank you to Kitini for giving me and everyone around you continuous joy and love for the past 10 years. As you watch me type this, I hope that I've finally made you proud.

Thank you to my mentors prior to arriving at Penn:

Dr. David Lederman at WVU for letting me join his lab, while still in high school, and for encouraging me to write up my findings for my first publication. That experience ensured I was hooked to research and science forever. Dr. Michael Clarke and Dr. Michael Rothenberg at Stanford for mentoring me during my first foray into biomedical research, where I discovered a love of stem cell biology. Dr. Hirdesh Uppal at Genentech for giving me the opportunity to grow as an independent scientist and enabling me to discover a love of bioengineering.

I have been exceptionally fortunate to have multiple incredible mentors during my PhD: Thank you to Dr. Christopher Lengner for serving as a clear example of both a great scientist and a kind person, everyday. Thank you for being supportive of my crazy ideas from the start and for giving me unparalleled freedom to pursue them. Thank you for our countless discussions, for responding to emails within minutes, and for always being available to strategize and give guidance. I sincerely hope that I have osmosed at least a few of your admirable qualities over the years.

Thank you to Dr. Jay Shendure and Dr. Aaron Mckenna for being incredibly gracious with your time over the last few years and providing invaluable advice with every comment during every meeting without fail.

Thank you to my thesis committee, particularly Dr. John Murray and Dr. Ben Stanger for guidance on computational techniques and pancreatic cancer biology, respectively.

Thank you to my present and past lab members for helpful discussions and for creating such a relaxed, supportive environment. Thank you to Dr. Jonathan Schug and Dr. Daniel Beiting for sequencing and computational resources and discussions. Thank you to Dr. Robert Norgard for always being willing to perform orthotopic transplants at a moment's notice. Finally, thank you to the Penn MSTP and especially to Dr. Skip Brass and Maggie Krall.



## ABSTRACT

### SINGLE-CELL LINEAGE TRACING OF CANCER METASTASIS

Kamen P. Simeonov

Christopher J. Lengner

The underpinnings of cancer metastasis remain poorly understood, in part due to a lack of tools for probing their emergence at high resolution. Here we present macsGESTALT, an inducible CRISPR-Cas9-based lineage recorder with highly efficient single-cell capture of both transcriptional and phylogenetic information. Applying macsGESTALT to a mouse model of metastatic pancreatic cancer, we recover ~380,000 CRISPR target sites and reconstruct dissemination of ~28,000 single cells across multiple metastatic sites. We find cells occupy a continuum of epithelial-to-mesenchymal transition (EMT) states. Metastatic potential peaks in rare, late-hybrid EMT states, which are aggressively selected from a predominately epithelial ancestral pool. The gene signatures of these late-hybrid EMT states are predictive of reduced survival in both human pancreatic and lung cancer patients, highlighting their relevance to clinical disease progression. Finally, we observe evidence for *in vivo* propagation of *S100* family gene expression across clonally distinct metastatic subpopulations.

<b>ACKNOWLEDGMENT .....</b>	<b>IV</b>
<b>ABSTRACT .....</b>	<b>VII</b>
<b>LIST OF TABLES.....</b>	<b>IX</b>
<b>LIST OF ILLUSTRATIONS .....</b>	<b>X</b>
<b>CHAPTER 1: ON CANCER METASTASIS AND LINEAGE TRACING .....</b>	<b>1</b>
SUCCESSES AND FAILURES IN CANCER TREATMENT AND SURVIVAL RATES .....	1
A GENETIC PREOCCUPATION AND A NON-GENETIC PROMISE .....	4
RETROSPECTIVE LINEAGE TRACING .....	6
PROSPECTIVE LINEAGE TRACING .....	11
<b>CHAPTER 2: SINGLE-CELL LINEAGE TRACING OF METASTATIC CANCER REVEALS SELECTION OF HYBRID EMT STATES .....</b>	<b>16</b>
INTRODUCTION.....	18
RESULTS .....	20
<i>An inducible lineage recorder with scRNA-seq readout .....</i>	<i>20</i>
<i>Aggressive clones are rare and transcriptionally divergent.....</i>	<i>21</i>
<i>An EMT continuum associated with aggression.....</i>	<i>25</i>
<i>Reconstruction of subclonal diversity arising in vivo.....</i>	<i>30</i>
<i>Late-hybrid EMT states are proliferatively and metastatically advantageous .....</i>	<i>33</i>
<i>Evidence for interclonal propagation of S100 gene expression .....</i>	<i>36</i>
DISCUSSION .....	40
ACKNOWLEDGEMENTS.....	42
AUTHOR CONTRIBUTIONS.....	43
DECLARATION OF INTERESTS.....	44
METHODS.....	45
<i>Key resources table.....</i>	<i>45</i>
<i>Resource availability.....</i>	<i>49</i>
<i>Experimental models and subject details .....</i>	<i>50</i>
<i>Method details.....</i>	<i>52</i>
<i>Quantification and statistical analysis .....</i>	<i>64</i>
ADDITIONAL RESOURCES.....	76
FIGURES .....	77
SUPPLEMENTARY FIGURES.....	86
SUPPLEMENTARY TABLES.....	98
<b>CHAPTER 3: DISCUSSION.....</b>	<b>99</b>
<b>BIBLIOGRAPHY .....</b>	<b>105</b>

## LIST OF TABLES

Resources table 1.....	45
Supplementary table 1 .....	98
Supplementary table 2 .....	98
Supplementary table 3 .....	98
Supplementary table 4 .....	98
Supplementary table 5 .....	98
Supplementary table 6 .....	98

## LIST OF ILLUSTRATIONS

Figure 1.....	77
Figure 2.....	78
Figure 3.....	80
Figure 4.....	81
Figure 5.....	83
Figure 6.....	84
Supplementary figure 1 .....	86
Supplementary figure 2 .....	89
Supplementary figure 3 .....	91
Supplementary figure 4 .....	93
Supplementary figure 5 .....	95
Supplementary figure 6 .....	96

## CHAPTER 1: ON CANCER METASTASIS AND LINEAGE TRACING

### Successes and failures in cancer treatment and survival rates

Cancer kills 10 million people annually. The vast majority of these deaths are due to metastasis — the process by which cancer cells spread from their origin and colonize distant tissues, thereby transforming a localized, often curable lesion into a systemic, largely incurable disease (Cancer Facts & Figures 2020; McGranahan and Swanton 2017).

The problem that cancer poses to humanity is only accelerating. As the world's population continues to age, with the median age rising from 21.5 to 30 years over the last 5 decades (Ritchie and Roser 2019), cancer burden continues to increase. In the United States, cancer care costs have rapidly increased annually, reaching nearly \$200 billion in 2020 (Cancer Action Network 2020). Contrasting this figure to the \$100 billion spent on cancer research by the National Cancer Institute (NCI), since the 1971 National Cancer Act, i.e. the start of the "War on Cancer", paints a dramatically dismal picture (Marshall 2011). Even more pessimistically, growth of the NCI's annual funding budget has not kept pace with inflation for the last 20 years, peaking in 2003 and 2009, but falling and stagnating since ("NCI Budget and Appropriations" 2015). However, these numbers do not account for private biotech and pharmaceutical spending on oncology research and development, as well as other cancer-focused government initiatives not falling under the umbrella of the NCI. Additionally, continued broad investment in basic research has provided both direct and indirect benefits to cancer research. A clear example is the Human Genome Project, which enabled The Cancer Genome Atlas

(TCGA), spurring a vastly improved understanding of the genetic underpinnings of cancer formation (Hutter and Zenklusen 2018).

By sequencing thousands of tumor samples, TCGA efforts have uncovered many recurrent genetic drivers of cancer formation (Kandoth et al. 2013). In some cancers, genetic drivers are incredibly common. For example, in non-hypermutated colorectal adenocarcinoma (COAD), the most common type of colorectal cancer, mutations in *APC*, *P53*, and *KRAS* are observed in 82%, 59%, and 45% of patients, respectively (Cancer Genome Atlas Network 2012). *P53* is recurrently mutated across many cancers, including, remarkably, in 95% of all ovarian cancers (OV) (Kandoth et al. 2013). In pancreatic ductal adenocarcinoma (PDAC), the most common type of pancreatic cancer, *KRAS* and *P53* are mutated in 93% and 72% of tumors, respectively (Cancer Genome Atlas Research Network 2017).

Critically, the identification of recurrent driver mutations has enabled research efforts to focus on the most functionally relevant genes and pathways for cancer formation and the discovery of potentially targeted therapeutics with low off-target toxicity. Indeed, in the last two decades, there has been a flurry of new targeted drug development, ranging from monoclonal antibodies, tyrosine kinase inhibitors, antibody-drug conjugates, checkpoint inhibitors, and engineered cell based therapies – over 100 targeted therapies for a variety of cancers have been approved (“Reflecting on 20 Years of Progress” 2021).

So, have survival rates of cancer improved since the start of the War on Cancer? At first glance, the 5-year survival rate for all cancers (non-benign) in the United States (US) has improved from 49% in 1975 to 69% in 2012 ([SEER 2017](#)). Unfortunately, upon closer examination, much of this improvement came before 2000, when the survival rate was already 66%. Furthermore, when stratifying by the stage of the cancer or degree of invasion, the small improvement from 2000-2012 appears to be primarily in patients with localized or regional disease at the time of diagnosis ([Esposito, Ganesan, and Kang 2021](#)). In fact, improved early detection is considered to be one of the primary reasons for the increased survival rates between the 1970s and the 2010s. As cancers are detected earlier and earlier, a larger proportion of cancers are detected at the localized disease stage, which has an excellent prognosis for many cancers. The 5-year survival rate for localized disease in breast cancer and melanoma is 99%, with colorectal cancer at 90% ([Cancer Facts & Figures 2020](#); [Esposito, Ganesan, and Kang 2021](#)).

Unfortunately, early detection is not always possible, and cancer is often already disseminated upon diagnosis. This is particularly apparent in PDAC, where 90% of patients have distant dissemination at the time of diagnosis ([Cancer Facts & Figures 2020](#)). Five-year survival data is dismal across all cancers when distant dissemination is present, some examples include: breast 27%, melanoma 25%, colorectal 14%, lung 5%, stomach 5%, pancreatic 3%, liver 2% ([Cancer Facts & Figures 2020](#)). These statistics are profoundly disturbing. Why despite all of the progress and targeted therapeutics over the last decades have survival rates for disseminated disease remained so astoundingly low?

## **A genetic preoccupation and a non-genetic promise**

Until recently, the majority of cancer research has focused on characterizing cancer genetics. And rightly so, genetic alterations have proven critical to transforming a normal tissue into an abnormal neoplasm. However, despite the successes outlined above of identifying stereotyped genetic drivers of cancer formation, recurrent genetic drivers of metastasis have proven elusive (Pereira et al. 2015; Margonis et al. 2015; M. K. H. Hong et al. 2015), with most cancers having no known drivers of metastasis (Makohon-Moore et al. 2017; Brastianos et al. 2015; McCreery et al. 2015; Yachida et al. 2010; Hunter et al. 2018).

As a result, non-genetic, i.e. epigenetic and transcriptional, changes have increasingly been purported to play a major role in driving metastasis (Hunter et al. 2018; Esposito, Ganesan, and Kang 2021). Interestingly, a similar trend has emerged in drug resistance research, where up to 40% of therapy resistant tumors fail to display any explanatory genetic adaptations for resistance (Marine, Dawson, and Dawson 2020). In the case of both metastasis and drug resistance, a series of interrelated, and potentially transcriptionally-driven, phenotypic changes have been implicated, including the cancer stem cell hypothesis (Todaro et al. 2007; Shackleton et al. 2009; Auffinger et al. 2014), epithelial-mesenchymal transition (EMT) (Yang et al. 2004, 2020; Fischer et al. 2015; Zheng et al. 2015), and cellular dedifferentiation and plasticity (Davis et al. 1986; Yang et al. 2020; Auffinger et al. 2014; Boumahdi and de Sauvage 2020; Liao et al. 2017).

However, the precise role that these processes play in metastasis, drug resistance, or both is actively under debate and poorly understood (Fischer et al. 2015; Zheng et al.



2015; Aiello et al. 2017; Quintana et al. 2008; Esposito, Ganesan, and Kang 2021; Ocaña et al. 2012; Tsai et al. 2012). Interestingly, these programs appear to be heavily interconnected and overlapping (Scheel and Weinberg 2012; Yang et al. 2020; Lambert and Weinberg 2021), often via the hijacking of primitive or developmental gene expression programs such as, Notch (Sethi et al. 2011; Domingo-Domenech et al. 2012; Takebe et al. 2011; Wu et al. 2017), Wnt (Esposito et al. 2019; Zhuang et al. 2017; DiMeo et al. 2009), TGF- $\beta$  (Calon et al. 2012; Padua et al. 2008; Colak and ten Dijke 2017), and Hedgehog (Yauch et al. 2008; Altaba and Ruiz i Altaba 2011). Notably, phenotypic and transcriptional adaptations, such as cancer "stemness" and EMT, are implicated across a variety of distinct cancer types, underscoring the immense therapeutic potential of building a better understanding of these processes in the context of cancer.

In order to build an understanding of how metastasis develops, we must obtain dense, precise, and accurate information of two major types: 1) the natural history of metastasis and 2) the molecular adaptations that vary along this natural history. The first of these would allow us to accurately identify which specific cancer subclones have the ability to metastasize, while the second tells us what alterations these subclones possess.

Another way to state these two types of information is: What is the clonal architecture or lineage of the cancer and the associated intratumoral heterogeneity or cellular qualities?

Until recently, the lineage tracing or reconstruction tools needed to attain such information have been lacking. They fall within two general categories – retrospective and prospective – in the next two sections we will examine both of these individually.

## **Retrospective lineage tracing**

In order to understand metastasis, we must reconstruct an accurate picture of the natural history of dissemination. In humans, where introduction of artificial lineage markers would be unethical or impossible, it is difficult to uncover the underlying population structure. Retrospective lineage tracing or reconstruction methods attempt to solve this problem by studying natural genetic variation in a population of cells or samples. Termed "retrospective" due to the analysis of genetic markers that have already occurred and did not occur by design.

As cells divide, a variety of error prone and stochastic processes are at work, resulting in genetic alterations which are heritable and serve to mark different lineages of proliferating cells that comprise a tissue or tumor. Errors accrue in the genome over the course of normal human development and homeostasis, as is illustrated by four recently copublished studies that characterized this genetic diversity in depth to explore both early developmental and adult tissue progenitor dynamics (S. Park et al. 2021; Coorens et al. 2021; R. Li et al. 2021; Moore et al. 2021; Naxerova 2021). This recent flurry of papers build on a continuously improving ability to characterize such mutations with greater depth and breath, paralleling advances in sequencing, and thereby revealing insights into normal human biology (Behjati et al. 2014; Ju et al. 2017).

A feature of the cancer genome is that it has a dramatically increased rate of accruing mutations and large-scale alterations (Stratton, Campbell, and Futreal 2009). This allows for cancer to rapidly adapt in its later stages, for example with the development of

resistance to a therapeutic, as is dramatically observed with BRAF inhibitors in melanoma ([Shi et al. 2014](#)). Interestingly, while increased genetic instability in cancer is often associated with a worse prognosis ([Carter et al. 2006](#); [Davoli et al. 2017](#)), too much instability may be counterproductive ([Birkbak et al. 2011](#); [Jamal-Hanjani et al. 2015](#)). For the purposes of studying cancer lineages, increased genetic instability and replication errors provide more substrate for the reconstruction of phylogenies than is often possible in normal tissues. Furthermore, unlike normal samples, tumor samples are often readily available to researchers due to surgical excision, particularly primary tumors.

The major sources of genetic variations that are employed as retrospective barcodes for lineage building in cancer are: Single-Nucleotide Variants (SNVs), Copy Number Variations (CNVs), microsatellites, and Long Interspersed Nuclear Element 1 (LINE-1) transposable retroelements. ([Baron and van Oudenaarden 2019](#)) provide a good brief overview of these types of variation in the context of lineage tracing more broadly, while ([Naxerova and Jain 2015](#)) provide a more focused and deep review of these methods in the context of studying cancer metastasis in human samples.

SNVs have been classically employed in population level studies via Genome Wide Association Studies (GWAS), but cancer's increased mutagenesis rate has also permitted their widely adopted use in intratumoral studies ([Hajirasouliha, Mahmoody, and Raphael 2014](#); [El-Kebir et al. 2015](#); [Popic et al. 2015](#)). A benefit of SNV-based analysis is that coding and driver mutations can be recovered alongside many passive

mutations used for lineage reconstruction. The major drawback is the rarity of these variations across the enormous morass of the genome (Schwartz and Schäffer 2017).

CNVs are large-scale, often chromosomal changes, where genomic regions that are greater than 1kb vary in their copy numbers (Baron and van Oudenaarden 2019). CNVs pose a benefit in that they are easier to measure with less sequencing depth and can often be inferred from even transcriptional data by analyzing differences in gene expression in the context of their genomic loci (Serin Harmanci, Harmanci, and Zhou 2020). CNVs have been used to build phylogenies from human tumor samples (Uchi et al. 2016; Ha et al. 2014).

Microsatellites are short stretches of repeated nucleotides that can trigger increased DNA polymerase skipping, which results in inappropriate loss or gain of nucleotides. Microsatellites are thus hypermutable regions of DNA, which are able to reach mutation rates 1000x higher than non-repetitive DNA and that have the added benefit of being at defined loci that can be examined via easier and cheaper targeted sequencing (Naxerova and Jain 2015). One study used targeted sequencing of polyguanine tracts to investigate hematogenous versus lymphatic spread in colorectal cancer (Naxerova et al. 2017).

Lineage reconstruction based on LINE-1 transposons works differently than the previously discussed approaches. LINE-1 elements are numerous throughout the genome and are able to undergo transposition (Ostertag and Kazazian 2001).

Sequencing of LINE-1 loci allows their genomic integration sites to serve as lineage markers.

While all of the above methods have various advantages and disadvantages when compared amongst each, they all suffer from three major, shared limitations. First, the lineage trees produced are quite coarse, with limited branches distinguishing subpopulations. Second, they often rely on bulk tissue samples to obtain enough sequencing depth in order to perform analyses. This imposes further limits on resolution and specifically creates the possibility of missing important clonal information that is either unsampled or masked due to rarity in sampled regions. Third, these analyses, focused on genetic variation for lineage purposes, also only obtain genetic information for cellular identity purposes, missing transcriptional or epigenetic information that may be critical to understanding drivers of metastasis, as discussed extensively in the previous section.

There have been some exceptions to these limitations. Landmark studies have used SNVs and CNVs to reconstruct single-cell level tumor lineages from 100 single cells ([Navin et al. 2011](#)), which has been scaled more recently to 1,300 single cells ([Casasent et al. 2018](#)). However, such approaches are challenging and expensive and as a result sequence vastly smaller numbers of tumor cells than what is now possible, at reasonable cost, with single-cell RNA sequencing (scRNA-seq) commercial assays. Moreover, they also do not capture potentially critical non-genetic information.

In the last two years, two methods have been reported that perform retrospective lineage tracing at the single-cell level at a massive-scale and concurrently capture transcriptional or epigenetic information via scRNA-seq or single-cell assay for transposase-accessible chromatin sequencing (scATAC-seq) [\(Nam et al. 2019; Ludwig et al. 2019\)](#). Genotyping of Transcriptomes (GoT) employed targeted sequencing of specific genomic loci of interest to build coarse clonal level information that is overlaid onto many thousands of single-cancer-cell transcriptomes [\(Nam et al. 2019\)](#). Meanwhile, [\(Ludwig et al. 2019\)](#) reported that mitochondrial DNA (mtDNA) is inadvertently captured by standard scRNA-seq and scATAC-seq, eventually optimizing this in [\(Lareau et al. 2021\)](#). mtDNA has an error rate that is 10-100 times higher than genomic DNA (gDNA), making it well-suited for lineage reconstruction [\(E. Kang et al. 2016; Biezuner et al. 2016\)](#). [\(Ludwig et al. 2019\)](#) and [\(Lareau et al. 2021\)](#) used this mtDNA combined with scRNA-seq and scATAC-seq to investigate clonality in normal hematopoiesis and leukemia.

It will be exciting to watch how these approaches will be used to vastly increase the scale of single-cell level retrospective lineage tracing studies that also annotate cells with non-genetic information. Especially, as the mtDNA based methodology appears to be rapidly advancing and is highly cost effective. However, it is important to note that even in these recent mtDNA based studies, while many thousands of cells are recovered, only a few dozen subclones are identified. Despite these recent advances, retrospective lineage tracing cannot supplant prospective lineage tracing methods, which themselves are advancing at breakneck pace. In the next section, we turn our attention to the recently invigorated field of prospective lineage tracing.

## Prospective lineage tracing

The value of prospective lineage tracing to study cancer dissemination has been recognized since the dawn of metastasis research. In a seminal study, (Talmadge, Wolman, and Fidler 1982) x-irradiated a melanoma cell line to induce chromosomal breakage. The resulting chromosomal rearrangements effectively served as lineage markers with which to trace how cells metastasized to the lung after implantation into the footpads of syngenic mice. Cells from metastases were isolated and cultured and subsequently karyotyped as a readout of the lineage label. They found that metastases were primarily monoclonal but also often originated from different clones in the primary tumor.

Since then a plethora of increasingly powerful prospective lineage tracing approaches have been developed, some of which have been used to study cancer metastasis. In this section, we will examine many of these technologies, particularly focusing on the recent proliferation of "evolving barcodes" or "lineage recording" systems.

A lineage tracer must be heritable, while not functionally altering its host cells. The first major evolution of prospective lineage tracing came with the application of green fluorescent protein (GFP) as a genetic marker in eukaryotic cells (Chalfie et al. 1994). Shortly thereafter, fluorescent proteins (FPs) were combined with genetic recombination, particularly *Cre-loxP* recombination in mammalian systems, to great effect. Tools using FPs whose expression was irreversibly activated via Cre and CreER systems driven by ubiquitous or cell-type specific promoters became a staple of studies on cell lineage (Kretzschmar and Watt 2012). Tools such as *Brainbow* and *Confetti* were developed that

combined multiple different FPs to produce a wider variety of stochastic outcomes in recombination and thus label multiple clones in the same starting population (Livet et al. 2007; Snippert et al. 2010).

FPs as standalone markers or in combination with recombinases continue to be used to study cancer metastasis, with recent studies shedding light on polyclonal metastatic seeding in various models (Aceto et al. 2014; Maddipati and Stanger 2015). However, while FP based lineage tracing is a powerful and widely-used tool to study not just metastasis, but also tumor clonality more broadly, it lacks labeling diversity and tracing resolution. Even systems that use multiple randomly recombined FPs to generate more color combinations are limited to labeling no more than ten clones, while mouse metastases and primary tumors can contain many thousands and millions of cells, respectively.

The next major evolution in prospective lineage tracing came with the advent of "lentiviral barcoding" or more generally "static barcoding", which integrates random or semi-random nucleotide sequences into the genomes of a starting population of cells, which usually receive the barcodes while in culture. The first application of this technology traced thousands of hematopoietic stem cells upon engraftment *in vivo* to study normal hematopoiesis (R. Lu et al. 2011).

Static barcoding has since been applied extensively to cancer, for example to study drug resistance (Bhang et al. 2015), population hierarchy (Lan et al. 2017), and metastasis (Echeverria et al. 2018). More recently, expressed static barcodes have been coupled



with scRNA-seq readout to allow efficient capture of clonal label and cell identity at the single-cell level to study normal hematopoiesis ([Weinreb et al. 2020](#)). However, two major limitations of "static barcoding" are that it is generally restricted to (1) creating lineage labels *in vitro* and (2) introducing labeling diversity at a single time point. Static barcoding is therefore unable to fully interrogate heterogeneity that may arise after the point of labeling, for example *in vivo* during dissemination and metastatic seeding.

The desire to address the limitations of static barcoding and understand cell lineage with greater resolution and physiological relevance has given rise to the "evolving barcoding" or "lineage recording" generation of methods reviewed in: ([Baron and van Oudenaarden 2019](#); [Kebschull and Zador 2018](#); [McKenna and Gagnon 2019](#); [Kester and van Oudenaarden 2018](#)). The goal of these methods is to achieve repeated or even continuous introduction of genetic labeling diversity *in vivo*, thereby addressing both of the major limitations of static barcoding. The mechanics of these evolving barcoding methods have varied widely. In a method akin to a prospective version of LINE-1 based retrospective lineage tracing, Sleeping Beauty transposons were used to enable labeling of native (non-transplant) hematopoiesis ([Sun et al. 2014](#); [Rodriguez-Fraticelli et al. 2018](#)). Contemporaneously, *Polylox* was applied to also study native hematopoiesis ([Pei et al. 2017](#)). *Polylox* used Cre to stochastically recombine an array of *loxP* flanked barcodes upon induction.

Starting in 2016, just prior to the start of the dissertation work presented here, the application of CRISPR-Cas9 to lineage tracing produced an explosion of creative new methods. Starting with GESTALT (genome editing of synthetic target arrays for lineage

tracing), CRISPR-Cas9 was employed to mutagenize compact, integrated lineage barcodes ([McKenna et al. 2016](#)). Inheritance patterns of barcode mutations could then be used to infer cellular phylogeny, akin to the role natural CNVs and SNVs serve at the whole genome level in retrospective lineage reconstruction. In the past five years, a flurry of new evolving barcoding approaches have been reported, primarily based around Cas9 mediated editing. These include the following methods, organized by their characteristics and major innovations — scRNA-seq readout of single barcode integrants: scGESTALT ([Raj, Gagnon, and Schier 2018](#)), LINNAEUS ([Spanjaard et al. 2018](#)), ScarTrace ([Alemany et al. 2018](#)), CARLIN ([Bowling et al. 2020](#)); *in situ* readout of barcodes: MEMOIR ([Frieda et al. 2017](#)) and intMEMOIR (a recombinase-based version) ([Chow et al. 2021](#)); self-targeting guide RNAs (gRNAs): homing barcodes ([Kalhor, Mali, and Church 2017](#); [Kalhor et al. 2018](#)), mSCRIBE ([Perli, Cui, and Lu 2016](#)), and CHYRON (insertion-biased homing barcodes) ([Loveless et al. 2021](#)); multiple expressed barcode integrations per cell: molecular recording ([Chan et al. 2019](#)).

The above methods were generally applied to study normal development in zebrafish or mouse, with one exception which focused on adult mouse hematopoiesis ([Bowling et al. 2020](#)). In 2021, three studies reported the use of evolving barcodes to study cancer metastasis ([Quinn et al. 2021](#); [W. Zhang et al. 2021](#); [Simeonov et al. 2021](#)). ([W. Zhang et al. 2021](#)) applied an existing method, homing barcodes ([Kalhor, Mali, and Church 2017](#); [Kalhor et al. 2018](#)), with bulk DNA sequencing readout to examine the effects of the bone microenvironment in influencing metastatic-ability. Meanwhile, ([Quinn et al. 2021](#)) and ([Simeonov et al. 2021](#)) reported new methods, which combined both static barcoding and evolving barcoding together with scRNA-seq readout of barcodes to

reconstruct metastasis at the single-cell level. [\(Quinn et al. 2021\)](#) improved on the "molecular recorder", originally described in [\(Chan et al. 2019\)](#), while [\(Simeonov et al. 2021\)](#) described multiplexed, activatable, clonal and subclonal GESTALT (macsGESTALT). In the next chapter, I will discuss the development of macsGESTALT, and its application to characterize cancer metastasis at high-resolution.

## CHAPTER 2: SINGLE-CELL LINEAGE TRACING OF METASTATIC CANCER REVEALS SELECTION OF HYBRID EMT STATES

Kamen P Simeonov<sup>1,2,\*</sup>, China N Byrns<sup>1,3</sup>, Megan L Clark<sup>4</sup>, Robert J Norgard<sup>5</sup>, Beth Martin<sup>6</sup>, Ben Z Stanger<sup>5,7,12</sup>, Aaron McKenna<sup>8,\*</sup>, Jay Shendure<sup>6,9,10,11,\*</sup>, Christopher J Lengner<sup>2,7,12,13,\*</sup>

<sup>1</sup>Medical Scientist Training Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

<sup>2</sup>Department of Biomedical Sciences, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA, USA

<sup>3</sup>Department of Biology, University of Pennsylvania, Philadelphia, PA, USA

<sup>4</sup>Department of Pathology & Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

<sup>5</sup>Abramson Family Cancer Research Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

<sup>6</sup>Department of Genome Sciences, University of Washington, Seattle, WA, USA

<sup>7</sup>Department of Cell & Developmental Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

<sup>8</sup>Department of Molecular & Systems Biology, Dartmouth Geisel School of Medicine, Lebanon, NH, USA

<sup>9</sup>Allen Discovery Center for Cell Lineage Tracing, Seattle, WA, USA

<sup>10</sup>Brotman Baty Institute for Precision Medicine, Seattle, WA, USA

<sup>11</sup>Howard Hughes Medical Institute, Seattle, WA, USA

<sup>12</sup>Institute for Regenerative Medicine, University of Pennsylvania, Philadelphia, PA, USA

<sup>13</sup>Lead Contact

\*Correspondence to: kamen.simeonov@gmail.com (K.P.S.); aaron.mckenna@dartmouth.edu (A.M.); shendure@uw.edu (J.S.); lengner@vet.upenn.edu (C.J.L.)

## Summary

The underpinnings of cancer metastasis remain poorly understood, in part due to a lack of tools for probing their emergence at high resolution. Here we present macsGESTALT, an inducible CRISPR-Cas9-based lineage recorder with highly efficient single-cell capture of both transcriptional and phylogenetic information. Applying macsGESTALT to a mouse model of metastatic pancreatic cancer, we recover ~380,000 CRISPR target sites and reconstruct dissemination of ~28,000 single cells across multiple metastatic sites. We find cells occupy a continuum of epithelial-to-mesenchymal transition (EMT) states. Metastatic potential peaks in rare, late-hybrid EMT states, which are aggressively selected from a predominately epithelial ancestral pool. The gene signatures of these late-hybrid EMT states are predictive of reduced survival in both human pancreatic and lung cancer patients, highlighting their relevance to clinical disease progression. Finally, we observe evidence for *in vivo* propagation of *S100* family gene expression across clonally distinct metastatic subpopulations.

## Keywords

lineage tracing; lineage reconstruction; CRISPR lineage tracing; evolving barcodes; barcoding; lineage recorder; molecular recorder; metastasis; cancer; phylogenetics; epithelial-to-mesenchymal transition; EMT; single-cell; scRNA-seq; S100

## Introduction

The vast majority of cancer deaths are due to metastasis, a process that transforms a localized, often curable lesion into a systemic, largely incurable disease ([Hunter et al. 2018](#); [Turajlic and Swanton 2016](#)). Recurrent genetic drivers of metastasis have proven elusive, suggesting that other levels of dysregulation may principally drive the phenomenon ([Hunter et al. 2018](#)). Phylogenetic histories of cancer progression in individual patients, e.g. based on analyses of copy number variation (CNV) or somatic mutation, can inform how the cells comprising metastases are related to the primary tumor, as well as to one another ([Naxerova and Jain 2015](#)). However, such methods are restricted to natural genetic diversity and additionally fail to concomitantly capture the molecular phenotype of each profiled cell, limiting what can be learned about the cellular programs that underlie the development and success of distinct metastatic clones. An alternative to retrospective phylogenetic approaches are traditional prospective lineage tracing methods, such as lentiviral barcoding, which involve tagging cells with unique DNA barcodes ([R. Lu et al. 2011](#)). However, such "static" barcoding strategies are generally restricted to introducing labeling diversity *in vitro* and at a single time point. Therefore, they are unable to capture critical *in vivo* processes, including any selection of intraclonal genetic or epigenetic heterogeneity emerging after the point of labeling.

Beginning with GESTALT (genome editing of synthetic target arrays for lineage tracing) ([McKenna et al. 2016](#)), a new paradigm for *in vivo* lineage tracing has emerged, employing CRISPR-Cas9 to progressively and stochastically mutagenize a compact, genomically-integrated barcode, thereby producing patterns of edits that can be used to reconstruct phylogenetic relationships amongst cells ([McKenna and Gagnon 2019](#)).

Such methods can be coupled to single-cell RNA sequencing (scRNA-seq) to explicitly relate cell lineage histories to transcriptional states ([Raj et al. 2018](#); [Spanjaard et al. 2018](#); [Chan et al. 2019](#)). Until recently, GESTALT and related methods have primarily been applied to early development, e.g. by injection of components into zygotes and subsequent profiling of edited barcodes and single cell transcriptomes from the resulting organism ([Bowling et al. 2020](#); [Quinn et al. 2021](#)). This strategy is fundamentally difficult to translate across biological systems as it requires specialized injection and titration. Furthermore, as components are neither integrated nor inducible, such systems are not amenable to longer-term or time-delayed studies in adult animals. However, with refinement, CRISPR-Cas9-based lineage tracers hold potential to be useful in contexts outside of early development, such as the study of somatic stem cell dynamics or cancer metastasis.

## Results

### ***An inducible lineage recorder with scRNA-seq readout***

To this end, we developed macsGESTALT (multiplexed, activatable, clonal and subclonal GESTALT), an integrated, inducible, and scalable method that can be easily adapted to any engineerable mammalian system to enable lineage tracing (**Figure 1**).

Our approach consists of three components (**Figure 1A**):

1) Each cell contains multiple unique barcode integrations. Barcodes are constitutively expressed within the 3' untranslated region (UTR) of a polyadenylated *pac* (puromycin *N*-acetyl-transferase) transcript, enabling sequencing via standard mRNA-based capture. Each barcode is a combination of a static 10bp sequence of random bases, used for clonal reconstruction, and a 250bp editable, evolving region composed of five CRISPR target sites, used for phylogenetic reconstruction (**Figures 1B-E**).

2) The evolving region is targeted by an array of five guide RNAs (gRNAs), separated by transfer RNA (tRNA) spacers, under a single constitutive mammalian U6 promoter. Upon transcription, tRNAs are excised from the array by endogenous RNase P and Z, releasing the individual gRNAs (Port and Bullock 2016). We selected this configuration from a screen of five different arrays, ranging from least compact to most compact (**Figures S1A-G**). The gRNA-tRNA array (**Figures S1E**) outperformed other compact configurations (**Figures S1F-G**) and similarly to the standard approach of placing each gRNA under its own U6 promoter (**Figures S1D**). Therefore, we selected the gRNA-tRNA configuration for its robust editing and compact size, allowing for easy transfer to



different vectors or promoters, consistent with our goals of creating an adaptable and broadly applicable system. These results also illustrate the usefulness of a tRNA spacing strategy for gRNA multiplexing in mammalian systems.

3) Cas9 expression and barcode editing are induced by doxycycline (dox) binding to a constitutive reverse tetracycline transactivator (rtTA) and activating a tetracycline responsive element (TRE) promoter ([Jian Cao et al. 2016](#)). Inducible barcode editing *in vitro* was robustly driven with limited leakiness, mostly confined to the first target site (**Figures S1H-K**). We also validated successful barcode recovery and clonal reconstruction in two independent experiments, each involving limiting dilution, expansion, and single cell sequencing (**Figures S1L-P**).

### ***Aggressive clones are rare and transcriptionally divergent***

We next set out to investigate cancer metastasis at high resolution by combining macsGESTALT and scRNA-seq ([Raj et al. 2018](#); [Chan et al. 2019](#)). We focused on pancreatic ductal adenocarcinoma (PDAC), which has a 5-year survival rate of 9%, the lowest of any major cancer ([Cancer Facts & Figures 2020](#)). Furthermore, 90% of PDAC patients have some dissemination at the time of diagnosis ([Cancer Facts & Figures 2020](#)). To study PDAC metastasis, we employed a commonly used model, where cells from KPCY (*LSL-Kras<sup>G12D</sup>; Trp53<sup>LSL-R172H</sup>; Pdx1-cre; LSL-Rosa26<sup>YFP/YFP</sup>*) mouse tumors ([Hingorani et al. 2005](#); [Rhim et al. 2012](#); [J. Li et al. 2018](#)) are orthotopically transplanted into the pancreata of non-tumor-bearing mice ([Rhim et al. 2012](#); [Aiello, Rhim, and Stanger 2016](#)). This approach presents highly consistent growth and metastasis kinetics and seeding patterns, and furthermore faithfully models human disease, due to the

following: 1) *Kras* gain-of-function and *p53* loss-of-function are the most common drivers of human PDAC ([Cancer Genome Atlas Research Network 2017](#)); 2) cells experience minimal time *in vitro* — a drawback of traditional cell lines; 3) a focal lesion develops in the pancreas that 4) disseminates to the same sites as human PDAC, including the liver and lung.

To investigate PDAC metastasis and associated transcriptional states, we selected a highly metastatic line from a library of characterized PDAC lines derived from KPCY tumors ([J. Li et al. 2018](#)) (**Methods**). To enable lineage tracing of these cells, we introduced dox-inducible Cas9 and the gRNA array through lentiviral transduction, and separately introduced multiplexed barcodes via PiggyBac-transposition, thereby producing macsGESTALT PDAC cells (**Figures 1D and 2A**). To model cancer metastasis *in vivo*, we injected mouse pancreata with 30,000 macsGESTALT PDAC cells, representing thousands of static barcode clones (**Figure 2A; Methods**). After one week of engraftment, we administered doxycycline in the drinking water to initiate lineage tracing. As expected, all mice were morbid at five weeks post-injection ([Aiello, Rhim, and Stanger 2016](#)). We randomly selected two mice, M1 and M2, and harvested cells from six cancer-bearing sites: primary tumor, liver, lung, peritoneal mets, surgical-site met (a peritoneal met forming at the peritoneal surgical incision site), and circulating tumor cells (**Methods**). PDAC cells were fluorescence sorted and processed for scRNA-seq of transcriptomes and macsGESTALT barcodes.

Overall, 89% of transcriptomes had corresponding clonal lineage information for M1 and 77% for M2, demonstrating improved barcode recovery using macsGESTALT compared

to prior methods ([Raj et al. 2018](#); [Bowling et al. 2020](#)). Notably, we observed a positive correlation between the recovery of a cell's transcriptomic RNA and barcode RNA ( $r = 0.64$ ,  $p < 2.2 \times 10^{-16}$ ) (**Figure S2A**). While the majority of cells had 10,000-100,000 transcriptome-derived transcripts and 10-100 barcode-derived transcripts, lower quality cells with low transcriptome recovery ( $< 5000$  transcripts), often had barcode recovery at the limit of detection (1-2 transcripts). Cells entirely lacking barcode information appeared to be a natural extension of this trend, as we recovered on average less than half of the overall transcriptomic RNA from these cells relative to those with barcodes recovered (Welch's t-test,  $p < 2.2 \times 10^{-16}$ ) (**Figure S2B**). Thus, barcode recovery appeared to be a function of cell quality and total RNA recovery rather than resulting from any specific bias or silencing event. With this in mind, we retained only cells with both high quality transcriptome and barcode information for downstream analyses (**Figures S2C-K**).

In total, across all sites in both mice, we recovered both the transcriptome and clonal history for 28,028 single cells (M1: 12,657; M2: 15,371) (**Figures S2C-K**). The set of static barcodes defining a clone were determined via hierarchical clustering and custom pipelines (**Methods**). Cells were then sorted into each clone based on their static barcode sequences, permitting even cells with missing barcodes to be assigned to the appropriate clone, while also enabling explicit multiplet detection and filtration and resulting in only ~0.5% unmatched cells (M1: 0.54% and M2: 0.51%) (**Figure S2J**). For M1, an average of 3.7 out of a possible 5.9 barcodes were recovered per cell, while recovery for M2 was on average 1.7 out of a possible 2.5 barcodes (**Figure S2J**). The

lower number of barcodes per cell in M2 likely contributed to its lower overall lineage recovery.

Clonal reconstruction revealed 95 distinct clones across the two mice (**Figure 2B**), identified by 227 static barcodes (**Figure S2J**), indicating that less than 1% of all injected clones successfully engraft. In contrast, *in vitro* experiments using the same cells and a similar time course revealed that most cells (clones) survive and form colonies on plates (**Figures S1L-P**). Thus, cancer cells in this model experience dramatic bottlenecks during *in vivo* engraftment.

Among the surviving clones, fitness differences were pronounced and shaped population structure across sites (**Figures 2B and 2C**). In the primary tumor, the majority (>50%) of cells came from a minority of clones (2 clones in M1; 6 clones in M2). Bottlenecking was even more extensive at metastatic sites, wherein 80-90% of cells typically came from a single clone (**Figures 2B and 2C**), and both mice had one clearly dominant clone across all disseminated sites (M1.1, M2.2). On the other hand, 51% of clones (48/95) failed to metastasize at all, suggesting that mutations in *Kras* and *p53* alone do not ensure metastatic success.

We next asked whether clones were transcriptionally distinct. Indeed, cells from the same clone clustered together in UMAP space (**Figure 2D**). This was true of both large and small clones (**Figures 2D-G**). Importantly, this finding extended to cells harvested from different sites, suggesting that cells retain their clonal transcriptional identity even after dissemination (**Figure S3A**). These stable transcriptional differences may result

from either epigenetic drift or large-scale copy number changes, the latter observed in our data (**Figure S3B**) and a hallmark of PDAC chromosomal instability ([Campbell et al. 2010](#)).

Finally, we asked whether or not differences in clonal behavior corresponded to transcriptional differences. While clones had distinct transcriptional identities, we found that many overlapped in UMAP space (**Figures 2D-G**). Furthermore, 81% of clones (77/95 across both mice) primarily resided in a single transcriptional cluster, Cluster 3 (**Figures 2B and 2H**). To relate transcriptional state to tumor aggression, we derived a clonal aggression scoring system based on clone size and dissemination (**Figure 2B; Methods**). We found that 85% (81/95) of clones were non-aggressive and were transcriptionally similar, occupying a small region of Cluster 3 (**Figures 2I and 2J**). Conversely, highly-aggressive clones were exceedingly rare but transcriptionally divergent from other clones and each other (**Figure 2I**).

### ***An EMT continuum associated with aggression***

We sought to understand the specific transcriptional programs associated with clonal aggression. While both mice were strikingly similar in terms of clonal composition (**Figure 2B**), we initially focused on M1, since we harvested cells from more sites and recovered over twice as many barcodes per cell, which permits more effective downstream subclonal reconstruction (**Figures S2J and S2K**). Reanalyzing the M1 data apart from M2, non-aggressive clones again appeared transcriptionally similar to one another (**Figure 3A**). Interestingly, these clones were enriched for expression of canonical epithelial markers, such as *Epcam*, *Muc1*, and *Cdh1* (**Figures 3B-D and**

**S4A**). Conversely, mesenchymal markers, such as *Sparc*, *Zeb2*, and *Col3a1*, were enriched in cells of the aggressive clone, M1.1 (**Figures 3E-G and S4B**). Loss of epithelial genes and gain of mesenchymal genes are defining hallmarks of epithelial-to-mesenchymal transition (EMT) (Nieto 2013; Nieto et al. 2016).

EMT is a process of transdifferentiation, wherein epithelial cells lose the properties of cell polarity and adhesion, while gaining the ability to be motile and migratory. In cancer, EMT is implicated in invasion, metastasis, tumor stemness, plasticity, and drug resistance (Nieto 2013; Nieto et al. 2016). EMT is primarily a transcriptional process mediated by a group of key master-regulator transcription factors (EMT-TFs) (Stemmler et al. 2019). We observed elevated expression in aggressive clones of 4/5 EMT-TFs, namely *Zeb1*, *Zeb2*, *Snai1*, and *Snai2* (**Figures 3F and S4C**). Expression of *Prrx1*, an important regulator of EMT in PDAC (Takano et al. 2016), was also increased.

Traditionally, EMT is considered a binary process, where cells switch from fully epithelial to fully mesenchymal. However, recent studies have reported discrete intermediate EMT states (M. Lu et al. 2013; J. Zhang et al. 2014; T. Hong et al. 2015; Pastushenko et al. 2018; Pastushenko and Blanpain 2019) or even a continuum of states (van Dijk et al. 2018; McFaline-Figueroa et al. 2019). In our data, epithelial and mesenchymal UMAP regions were not well segregated. Specifically, epithelial and mesenchymal genes appeared to gradually lose and gain expression as a function of distance from two extremes (**Figures 3B-G**), supporting the view that a continuum of EMT states exists *in vivo*.

We leveraged our single-cell data to explore the transcriptional correlates of EMT as a continuum. We performed unbiased trajectory inference using Monocle 3 ([Junyue Cao et al. 2019](#)) and found that the main trajectory in our data corresponded to the observed EMT gene expression axis (**Figure 3H**). We named this trajectory "pseudoEMT" (akin to pseudotime for developmental trajectories) and placed the root of the trajectory, or the zero EMT state, at the most epithelial transcriptional region (**Figure 3H**). Hence, the expression of canonical epithelial markers was highest at the root. We found that many genes, including known epithelial or mesenchymal markers, rise and fall at different rates across pseudoEMT (**Figures 3I and S4E-G**); for example, many extracellular matrix genes activate only very late in the trajectory (**Figures 3I and S4F**). Additionally, numerous genes, such as *Cd44* or *Inhba*, displayed unusual patterns, rising and then falling or plateauing (**Figure S4H**). Expression of surface markers previously used to stratify different EMT states in skin and breast cancer mouse models, *Epcam*, *Vcam1* (CD106), *Itgav* (CD51), and *Itgb3* (CD61) ([Pastushenko et al. 2018](#)), followed a similar pattern in our data (**Figures S4D**). However, except for *Epcam*, expression of these markers was not highly variable across the EMT continuum (**Figures S4I**), suggesting that at least in PDAC, other genes might be more suitable markers for stratification.

Plotting cells along pseudoEMT highlighted that smaller, non-aggressive clones reside on the epithelial extreme, while more mesenchymal states are restricted to large, aggressive clones, such as M1.2 and particularly M1.1 (**Figures 3I**). As 27 of 29 clones were highly epithelial, we suspected this to be the default transcriptional state. To investigate this, we applied single-cell RNA-seq on 5,932 *in vitro* cultured cells. We found that these cells comprised 40 distinct clones, none of which overlapped with any

clones recovered from *in vivo* metastasis experiments. *In vitro* cells clustered homogeneously together and away from M1 cells (**Figures S5A and S5B**) and had distinct markers from *in vivo* cells at large (**Figures S5C and S5G and Table S1**). With regards to EMT, *in vitro* cells were strikingly epithelial, often displaying higher expression of epithelial markers, such as *Muc1* and various keratins (**Figures S5D, S5E, and S5H**), and conversely even lower expression of mesenchymal markers, such as *Zeb2*, *Vim*, and *Fn1* (**Figures S5F and S5I**), as compared to the highly-epithelial clones of M1. Thus, the baseline state of these PDAC cells appears to be highly epithelial with more mesenchymal EMT states only appearing *in vivo*, as in M1.1 and M1.2.

To systematically characterize gene expression along EMT *in vivo*, we identified the top 3000 significantly differentially expressed genes across pseudoEMT ( $q \sim 0$ , Moran's  $I > 0.1$ ) (**Table S2**). Hierarchical clustering of genes revealed six gene sets with similar kinetics (**Figure 3J**). We classified these sets from most epithelial to most mesenchymal as follows: Epithelial (E), Hybrid 1, 2, 3, and 4 (H1, H2, H3, H4), and Mesenchymal (M) (**Figure 3J; Table S2**). We then performed hypergeometric gene set enrichment using the Molecular Signatures Database (MSigDB) Hallmark gene sets, which represent well-defined biological states and processes (**Figure 3J; Table S2**). Concordant with the pseudoEMT trajectory, gene set enrichment indicated an EMT process. Early clusters (E, H1) were enriched for apical surface genes, consistent with epithelial cell polarity, while late clusters showed gradually increased enrichment for EMT (H4:  $p = 3 \times 10^{-6}$ , M:  $p = 3 \times 10^{-29}$ ). An inducer of EMT and metastasis, TGF- $\beta$  signaling (Zavadil and Böttinger 2005; Nieto et al. 2016; Aiello et al. 2018), as well as Jak/Stat3 and Stat5 signaling (R.-Y. Liu et al. 2014), peaked in the late hybrid state (H4) and tapered off in the highly



mesenchymal state (M). Other pathways purported to be involved in EMT, such as TNF- $\alpha$  (Wang et al. 2013), Wnt (Kim, Lu, and Hay 2002; Basu, Cheriya-mundath, and Ben-Ze'ev 2018), and Hedgehog (J. Zhang, Tian, and Xing 2016) were also only enriched in H4 or M. Interestingly, Notch signaling was recently implicated as a hybrid-EMT stabilizer (Boareto et al. 2016; Bocci et al. 2017), consistent with our finding that it was only enriched in H4.

Striking metabolic gene expression changes across EMT were also apparent (**Figure 3J**). Transitioning from early (H1, H2) to late (H3, H4) hybrid gene clusters, we observed a strong shift from enrichment of oxidative phosphorylation (OXPHOS) toward glycolysis, potentially related to the enrichment of mTOR signaling in H2 (Ramanathan and Schreiber 2009). Consistent with metabolic shifts, hybrid EMT states also were highly enriched for proliferative gene sets, such as G2M, E2F, and mitotic spindle. Specifically, enrichment began modestly in H2 and peaked dramatically in H3 (G2M, H2:  $p = 3 \times 10^{-2}$ , H3:  $p = 1 \times 10^{-30}$ ). We next determined the cell cycle phase of each cell (G1, G2M, or S) to estimate the proportion of actively dividing cells (S/G2M) across pseudoEMT (**Methods**). Consistent with Hallmark gene set enrichment, cell cycling peaked at EMT regions representing the E and H2/H3 gene clusters (**Figure S4J**). These hybrid EMT proliferative changes were potentially driven by Myc (Gabay, Li, and Felsher 2014), as Myc targets mirrored proliferative gene set enrichment and cell cycling fraction (Myc-v1, H2:  $p = 1 \times 10^{-3}$ , H3:  $p = 1 \times 10^{-30}$ ).

We next asked which TFs might regulate progression through EMT. Applying HOMER (Heinz et al. 2010) to promoters, we detected 45 significantly enriched DNA motifs

binding factors across all gene clusters (**Figure 3K**). EMT master regulators, *Zeb1*, *Zeb2*, *Snai1*, and *Snai2*, were enriched in early clusters, E and H1. As EMT-TFs are primarily transcriptional repressors that downregulate epithelial genes (Stemmler et al. 2019), this finding illustrates our ability to discover regulators of the EMT continuum. ETS-domain TFs, which are associated with metastasis, invasion, and EMT (Hsu, Trojanowska, and Watson 2004; Sizemore et al. 2017), dominated the enrichment profiles of hybrid states H2 and H3. Motifs bound by members of the Sox and Fox families were enriched in H4 and M, respectively. Sox TFs are often associated with stemness-related processes (Grimm et al. 2019). Notably, the six gene clusters have no overlapping genes, yet adjacent clusters often displayed overlapping TF and gene set enrichment, lending further support for a gradual continuum of EMT transitions (**Figures 3J and 3K**). Overall, across this continuum of 3000 genes, we describe many classic EMT markers, pathways, and regulators, but we also find many less well-characterized genes and processes of potential interest for furthering understanding of EMT *in vivo* (**Table S2**). Additionally, we performed a traditional Leiden clustering of M1 and found clusters roughly matching the pseudoEMT spectrum (**Figure S5J**). We identified the top markers by both cluster and clone, finding that cluster markers were consistent with genes enriched across corresponding EMT states (**Table S3**).

### ***Reconstruction of subclonal diversity arising in vivo***

Most cells in the mid-to-late EMT continuum came from a single dominant clone, M1.1, preventing us from precisely correlating transcriptional processes with tumor aggression and highlighting the limitations of static barcoding (**Figure 3I**). We therefore leveraged

editing patterns of macGESTALT evolving barcodes to more precisely relate EMT and aggression at the subclonal level.

We recovered a large number of edited and informative target sites per cell, conducive to phylogenetic analysis. Altogether, we recovered 384,870 CRISPR target sites, of which 96% were edited (**Figure S6A**). Editing was distributed across the length of the barcodes with peaks at the expected Cas9 cut-sites, 3bp upstream of the protospacer adjacent motif (PAM) of each target site (**Figure 4A**). Deletions predominated over insertions, as expected ([McKenna et al. 2016](#); [Raj et al. 2018](#); [Bowling et al. 2020](#)), with an approximately equal number of single- and multi-target deletions (**Figures 4B and S6B**). The average edit size varied by edit type, with 11bp for insertions, 18bp for single-target deletions, and 80bp for multi-target deletions (**Figure S6C**). Multi-target deletions were of a large size range and involved 2, 3, 4, or 5 target sites at frequencies ranging from 10-19% (**Figures S6B and S6C**). Individual target site editing rates varied between 89-99% (**Figure 4B**). On average, we recovered 18.5 target sites (3.7 barcodes) per cell for M1 and 8.5 (1.7) for M2 (**Figure S2J**).

Intraclonal tree reconstruction was performed in three main steps (**Figure 4C**). First, different barcodes from the same cell were concatenated based on their static barcodes into a "barcode-of-barcodes", which contains all of the phylogenetic information recovered for that cell. Second, cells with identically edited barcode-of-barcodes were grouped into subclones, since they are indistinguishably close relatives. Third, phylogenetic relationships between subclones were reconstructed based on edit inheritance patterns (**Figure 4C**). Subclonal metastatic aggression was quantified via

Shannon's Equitability ( $E_H$ ) – a statistical measure of dissemination across harvest sites (**Methods**). For example, a subclone found at only one harvest site is not metastatically aggressive and has an  $E_H$  of zero.

We sought to understand the maximum number of cells that could be uniquely tagged using our approach. With this in mind, we first investigated editing diversity of individual barcode integrants (**Figures S6D**). Examining 208 barcodes across both mice, we found that the maximum number of unique editing outcomes for a barcode scaled with the number of cells recovered, but gradually peaked to around 400 unique outcomes even for barcodes recovered in nearly 10,000 cells. Hence, in these experiments where we recovered an average of 2.6 barcodes per cell, we can estimate maximum labeling at nearly  $10^6$  cells (400 editing outcomes  $\wedge$  2.6 barcodes \* 95 clones).

In practice, we sampled a fraction of this theoretical space and recovered 6,055 unique barcodes-of-barcodes, which for efficient phylogenetic reconstruction, we filtered to a total of 1,692 subclones, each with at least two cells for larger clones ( $\geq 50$  cells) or with any number of cells for smaller clones (**Figure S6A; Methods**). Due to a higher average number of barcode integrations per cell, M1 displayed greater reconstructive power than M2. This was particularly apparent in the dominant clone of each mouse, where M1.1 with seven barcode integrants had 601 subclones compared to M2.2 with only two integrants and 110 resulting subclones. Notably, pairwise phylogenetic distances in the reconstructed trees were strongly concordant with the corresponding edit distances between barcode-of-barcodes alleles (**Figure S6E**) and more active target sites

determined earlier tree nodes (**Figure S6F**), suggesting that lineage relationships between cells are accurately captured in our trees.

The full clonal and subclonal phylogenetic visualization of M1 data highlights the overwhelming proliferative and metastatic dominance of clone M1.1 (**Figure 4D and S6G**). However, within M1.1, we also observed vast heterogeneity with respect to subclonal aggression and metastatic success. Most strikingly, the same bottlenecks observed on the clonal level was also present on the subclonal level within M1.1 (**Figure 4E**). Subclonal bottlenecks further increased at metastatic sites, again mirroring observations at the clonal level. Thus, cancer progression appears to be defined by a state of constant selection, separate from the effects of engraftment.

***Late-hybrid EMT states are proliferatively and metastatically advantageous***

As the vast majority of EMT diversity was within M1.1 (**Figure 3I**), we leveraged phylogenetic data to understand how this range of intraclonal EMT states may relate to differences in subclonal behavior. We calculated the mean pseudoEMT value for each subclone and plotted this and subclonal dissemination ( $E_{ij}$ ) for clone M1.1 (**Figures 5A and 5B**). While M1.1 was highly mesenchymal compared to other M1 clones, many subclones within M1.1 were actually quite epithelial. These epithelial subclones were primarily small and non-metastatic (**Figures 5A and 5B**). Interestingly, the same was true of highly mesenchymal subclones. On the other hand, the largest and most disseminated subclones appeared to express hybrid EMT states (**Figures 5A and 5B**), providing direct evidence that EMT extremes are less metastatic than hybrid states ([Jolly](#)

et al. 2015; Nieto et al. 2016; Lambert, Pattabiraman, and Weinberg 2017; Pastushenko and Blanpain 2019).

To precisely characterize where aggression peaked along the EMT continuum, we mapped subclonal dissemination ( $E_{sc}$ ) and size along pseudoEMT (**Figure 5C**). We found that dissemination gradually peaked around the H3 and H4 hybrid states (pseudoEMT score of 20-22) and then sharply declined at highly mesenchymal states. Thus, late-hybrid EMT states are metastatically advantageous and are associated with specific proliferative, metabolic, and signaling processes (**Figure 3J and Table S2**), as well as distinct regulatory binding factors (**Figure 3K**).

Notably, hybrid-EMT states appeared transcriptionally stable – for example, a large, hybrid subclone often had close relatives that were also large and hybrid (**Figure 5A**). To understand the stability of EMT states, we plotted the distribution of cells, subclones, and root clades along pseudoEMT (**Figure 5D; Methods**). Root clades mark the first phylogenetic subdivision within a clone and are hence an older subgrouping of cells than a subclone. Examples of root clades and subclones are highlighted in **Figure 5A**. Root clades exist at the time of dox initiation (one week post orthotopic transplant), cells exist at the time of harvest, and subclones in between; thereby we compared different "levels" of ancestral groups. Moving from root clades to cells, there was a shift from epithelial to hybrid states, suggesting that while epithelial states are the prevailing ancestral default, they are proliferatively and metastatically disadvantaged compared to hybrid states (**Figure 5D**). This *intraclonal* observation again mirrored findings at the *interclonal* level, where M1.1 itself was dominant compared to all other clones, which were generally,

highly epithelial. Therefore, ongoing natural selection of rare, late-hybrid EMT states over predominating epithelial states permits both rapid dissemination and forces continuous clonal and subclonal bottlenecking.

As late-hybrid EMT states, namely the H3 and H4 gene clusters, were profoundly associated with metastasis in our model, we asked whether a similar trend might exist in human PDAC (**Figure 5E**). Using The Cancer Genome Atlas (TCGA) matched gene expression and clinical data, we found that the transcriptional signature of the E, H1, and H2 gene clusters had no association with disease prognosis. However, patients enriched for the H3 or H4 transcriptional signature had a significantly increased risk of death, and this risk disappeared for the highly mesenchymal cluster M (**Figure 5E**). Remarkably, these human PDAC findings faithfully mirror the rise and fall of subclonal metastatic aggression along pseudoEMT in our model (**Figure 5C**).

As EMT is thought to play a role across many cancer types ([Nieto et al. 2016](#)), we also examined whether our pseudoEMT gene sets might predict survival in the other prevalent cancers by mortality ([Cancer Facts & Figures 2020](#)): lung, colorectal, breast, and prostate cancer. While colorectal, breast, and prostate cancers were not significantly associated in either direction with our PDAC-derived pseudoEMT gene sets, lung cancer displayed a similar pattern to PDAC (**Table S4**). Lung cancer patients enriched for H4 had significantly worse overall survival, while those enriched for M again trended toward better overall survival. In summary, these findings highlight the clinical relevance of late-hybrid states and emphasize the potential cancer-specific nature of EMT.

### ***Evidence for interclonal propagation of S100 gene expression***

We also examined the lineage and transcriptional structure of M2, which overall appeared strikingly similar to M1 (**Figures 6A-B versus 3A and S6G**). As in M1, labeling the transcriptional UMAP of M2 by clone highlighted that non-aggressive clones occupy a similar transcriptional region, while rare metastatic clones and one dominant clone occupy divergent transcriptional regions (**Figure 6B, S5K and Table S3**).

However, due to the lower number of barcode integrants in M2.2 relative to M1.1 and the resulting lower number of subclones reconstructed (**Figure 6A versus S6G**), we were unable to interrogate the dominant clone of M2 in the same depth as M1.1. We instead broadly asked what genes might be associated with subclonal dissemination ( $E_{\text{H}}$ ) in M2, by performing a regression of  $E_{\text{H}}$  against single cell gene expression with adjustment for confounders (**Methods**). We identified 973 genes positively associated with dissemination and 1,037 negatively associated genes ( $q < 0.05$ ) (**Table S5**).

Promisingly, as in M1, genes positively associated with subclonal dissemination in M2 also predicted worse overall survival in human PDAC TCGA data (**Figure 6C**), as well as in human lung cancer but not in breast, colorectal, and prostate cancer (**Table S5**).

Meanwhile, amongst the genes most negatively associated with dissemination were canonical epithelial markers, such as *Ocln*, *Epcam*, and *Lgals4* (**Table S5**). These epithelial genes presented similar patterns of expression to as seen in M1. Adhesion encoding genes, *Ocln* and *Epcam*, were strictly contained to non-aggressive UMAP regions in M2 (**Figures 6D and 6E**), as they were in M1 (**Figures 3B and S4A**), while *Lgals4* was expressed slightly more broadly, just it was in M1 (**Figures 6F and S5E**). Thus, the vast majority of clones in both M1 and M2 were non-metastatic and epithelial



in nature. This finding, together with our observation that these cells express epithelial but not mesenchymal markers *in vitro* (**Figures S5D-F and S5H-I**), further indicates that the default state is epithelial, that epithelial markers are repressed in order to metastasize, and that this process is rare.

As in M1, EMT-TFs, *Prrx1* and *Zeb2*, were expressed inversely to epithelial genes (**Figures 6G and 6H**). However, while most aggressive clones in M2 displayed similar expression patterns to M1 with regards to epithelial and mesenchymal genes, the dominant clone, M2.2, was not entirely consistent with the canonical EMT axis observed in M1 (**Figure 3J**). Specifically, the mesenchymal marker, *Sparc*, was lowly expressed in non-aggressive regions but also in M2.2 (**Figure 6I**). Similarly, the epithelial marker *Muc1* was highly expressed both in non-aggressive regions and in a large portion of M2.2 cells (**Figure 6J**). This was particularly apparent when comparing M2.2 to another aggressive clone, M2.23 (**Figures 2B and 6B**), which displayed more canonical and complete EMT, with high mesenchymal gene expression (**Figures 6G-I**) and nearly completely absent epithelial gene expression (**Figures 6D-F and 6J**). Indeed when plotted together with M1, M2.23 cells clustered with the more mesenchymal cells of M1.1 (**Figure 2D**), which may help explain its aggressive but non-dominant phenotype (**Figure 2B; Methods**).

We sought to better understand the processes that underlie dominance of M2.2 and aggression in M2 more broadly. Thus, we narrowed the genes significantly associated with subclonal dissemination to those that were both highly expressed and had a strong association, leaving 355 genes (**Figure 6K; Methods**). Among the most negatively

associated genes were again epithelial markers, as well as genes such as *Ctse*, which has been functionally shown to inhibit tumor growth and metastasis ([Kawakubo et al. 2007](#)). Conversely, among the most positively associated genes were genes previously found to promote TGF- $\beta$  signaling, EMT, and metastasis in other cancers, such as *Ifitm1*, *Ifitm3*, and *Akr1b3*, further highlighting the important role EMT plays in promoting metastasis across both M1 and M2 ([Yu et al. 2015](#); [X. Liu et al. 2019](#); [Min et al. 2018](#); [Schwab et al. 2018](#)).

Notably, we found that the *S100a* gene family was 52-fold over-enriched among positively associated genes (hypergeometric test,  $p = 8 \times 10^{-10}$ ) and completely absent from negatively associated genes (**Figure 6K**). S100 proteins were recently found to be the most abundant and overrepresented secreted factors in PDAC compared to normal pancreas, in both human patients and mouse models ([Tian et al. 2019](#)). However, the specific functions of S100s in PDAC and other cancers are poorly characterized. Some S100s, such as S100a4, are thought to promote metastasis via EMT and to directly mediate pseudopodia and lamellipodia formation in order to drive cell migration and invasion ([Bresnick, Weber, and Zimmer 2015](#); [Fei et al. 2017](#)). Interestingly, S100s are considered autocrine, paracrine, and even circulatory, long distance signaling molecules that potentially propagate their own expression and coordinate changes in the tumor and the microenvironment both locally and systemically ([Bresnick, Weber, and Zimmer 2015](#)). However, studies have primarily focused on S100 signaling in the tumor microenvironment and have not assessed how signaling spreads across different tumor subpopulations.

We leveraged our coupled lineage and transcriptional data across 95 distinct cancer clones to investigate whether there was evidence of *S100* signal propagation in tumors *in vivo*. We aggregated single-cell gene expression of the *S100a* family for each clone grouped by mouse (**Figure 6L**). We found that M2 clones had significantly higher expression of *S100a* genes compared to M1 clones (Welch's t-test,  $p = 9 \times 10^{-6}$ ) and that this was also true when restricting comparison to only the aggressive clones of each mouse ( $p = 2 \times 10^{-3}$ ). Notably, each of the 7 aggressive clones of M2 had higher *S100* expression than any of the 29 clones of M1 (**Figure 6L**). As all clones from both mice derive from the same starting population *in vitro* and are largely unrelated with unique histories, as evidenced by their macsGESTALT static barcodes (**Figure 2B**) as well as their distinct CNVs (**Figure S3B**), these findings present clear evidence of *S100* expression propagation across distinct clonal tumor populations *in vivo*. Furthermore, aggressive clones in M2 had significantly higher *S100* expression than non-aggressive clones ( $p = 6 \times 10^{-4}$ ), while this was not the case for M1 (**Figure 6L**). Indeed, M2.2, the dominant clone of M2, which displayed inconsistencies with regards to some canonical epithelial and mesenchymal markers, had the highest *S100a* expression of any clone across either mouse, suggesting that it had achieved dominance by complementing canonical EMT changes with high *S100* expression.

## Discussion

To study cancer metastasis at high resolution, we developed macsGESTALT, a multiplexed, inducible lineage tracer that can be easily coupled with scRNA-seq. We applied macsGESTALT to an *in vivo* model of pancreatic cancer metastasis and reconstructed transcriptomic information, lineage history, and harvest site for ~28,000 single cells derived from nearly 100 clones. These richly annotated cancer metastasis phylogenies can be explored interactively at <https://macsgestalt.mckennalab.org/>.

Despite extensive investigation, the identification of recurrent genetic drivers of metastasis has remained challenging (Hunter et al. 2018). Here, in spite of using a metastatically competent genetic model, we found that most clones in fact do not metastasize, supporting the importance of transcriptional and non-genetic processes in metastasis, such as acquisition of late-hybrid EMT states or propagation of *S100* expression. While our approach enabled us to precisely map the association between metastasis and EMT and thereby identify gene sets predictive of human survival, further functional investigation of specific EMT states is necessary (Zheng et al. 2015; Aiello et al. 2017, 2018). Similarly, the *S100* gene family appears to play a number of important yet poorly understood roles in cancer (Bresnick, Weber, and Zimmer 2015; Tian et al. 2019) and warrants further functional dissection of its many distinct family members. Additionally, direct comparison of our data to scRNA-seq from human patients may shed further light on the relevance of our findings to human disease.

In this study, we apply macsGESTALT lineage tracing to ~100 clones across two mice and find both conserved and distinct ways in which metastasis is achieved. We

anticipate that future studies will build on this work and exhaustively explore the full landscape of possible paths to metastasis. macsGESTALT is well suited for such a task, as its inducibility allows lineage tracing to initiate at the optimal experimental time, here after tumor engraftment. Alternatively, initiation can be coupled with specific interventions, such as the administration of a therapeutic to study chemoresistance. Future optimization of macsGESTALT may include editing rate titration, minimization of multi-target deletions, and coupling to other emerging technologies such as signal recording. These technical advancements will enable questions in cancer and stem cell biology to be investigated at previously inaccessible levels of resolution and scale.

## **Acknowledgements**

We thank J.I. Murray for advice on lineage and transcriptional analyses, J. Li for donation of PDAC cell line and advice on its use, J.A. Gagnon for advice on barcode editing and lineage tracing, and M.A. Blanco for advice on TCGA survival analysis. We thank K. Tan and A. Raj as well as all members of the Lengner laboratory for helpful discussions. We also thank the University of Pennsylvania Next-Generation Sequencing Core, in particular J. Schug and J. Kutch, for advice on barcode sequencing. We thank D.P. Beiting for computational resources. This research was supported by the Ruth L. Kirschstein National Research Service Award F30-DK120135, the Blavatnik Family Fellowship in Biomedical Research, and T32-HD083185 (to K.P.S.), National Human Genome Research Institute R00HG010152 and National Cancer Institute 5P30CA023108-37 (to A.M.), the Howard Hughes Medical Institute and Allen Discovery Center for Cell Lineage Tracing (to J.S.), and National Cancer Institute R01-CA168654 and the Shipley Foundation Program for Innovation in Stem Cell Science (to C.J.L.).

### **Author Contributions**

K.P.S. initiated, designed, and coordinated the study with the guidance of A.M., J.S., and C.J.L.; K.P.S., B.M., and A.M. constructed vectors; K.P.S. generated cell lines and performed in vitro experiments; R.J.N. performed orthotopic injections and advised on the PDAC model with B.Z.S.; K.P.S. and M.L.C. harvested and isolated tumor cells; K.P.S. performed bulk and single cell library preparation and sequencing, wrote clonal reconstruction and subclonal analysis scripts, and performed all single cell lineage and transcriptional analyses; C.N.B. performed motif enrichment, copy-number variation, and survival analyses; K.P.S. wrote the manuscript and generated all figures and data visualizations; K.P.S., C.N.B., M.L.C., A.M., J.S., and C.J.L. reviewed and edited the manuscript.

**Declaration of Interests**

The authors declare no competing interests.



## Methods

### Key resources table

Resources table 1

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
DMEM, High Glucose	Fisher Scientific	Cat#: 11-965-092
FBS	Corning	Cat#: 35-010-CV
L-Glutamine	Invitrogen	Cat#: 25030081
Penicillin-Streptomycin	Invitrogen	Cat#: 15140122
TrypLE Express Enzyme	Thermo Fisher Scientific	Cat#: 12605010
Collagenase IV	Thermo Fisher Scientific	Cat#: 17104019
Lipofectamine 3000	Thermo Fisher Scientific	Cat#: L3000001
Lipofectamine 2000	Thermo Fisher Scientific	Cat#: 11668030
Lipofectamine CRISPRMax	Thermo Fisher Scientific	Cat#: CMAX00001
G418	Invitrogen	Cat#: 108321-42-2
Puromycin	Sigma-Aldrich	Cat#: P8833
Doxycycline Hyclate	Sigma-Aldrich	Cat#: D9891
BSA	Sigma-Aldrich	Cat#: A7906
DAPI	Thermo Fisher Scientific	Cat#: 62248
EDTA	Invitrogen	Cat#: 15575020
DNase I	Sigma-Aldrich	Cat#: D4263
ACK Lysing Buffer	Quality Biological	Cat#: 118-156-721
HBSS	Invitrogen	Cat#: 14175079
PBS	Invitrogen	Cat#: MT21-031-CM
Critical commercial assays		
NEB Stable Competent E. coli	NEB	Cat#: 3040H
NEBuilder HiFi DNA Assembly Master Mix	NEB	Cat#: E2621
GeneArt Precision gRNA Synthesis Kit	Thermo Fisher Scientific	Cat#: A29377
NucleoSpin DNA RapidLyse Kit	Macherey-Nagel	Cat#: 740100.50
Agencourt AMPure XP	Beckman Coulter	Cat#: A63880
SPRI Select	Beckman Coulter	Cat#: B23317
TapeStation High Sensitivity D1000 ScreenTape	Agilent	Cat#: 5067-5584
TapeStation High Sensitivity D1000 Reagents	Agilent	Cat#: 5067-5585
TapeStation High Sensitivity D5000 ScreenTape	Agilent	Cat#: 5067-5592
TapeStation High Sensitivity D5000 Reagents	Agilent	Cat#: 5067-5593
Qubit 1X dsDNA HS Assay Kit	Thermo Fisher Scientific	Cat#: Q33230
NEBNext Multiplex Oligos for Illumina (Dual Index Primers Set)	NEB	Cat#: E7600S
HotStart ReadyMix	Kapa Biosystems	Cat#: KK2601

KAPA Real-Time Library Amplification Kit	Kapa Biosystems	Cat#: KK2702
MiSeq Reagent Kit v3 (600-cycle)	Illumina	Cat#: MS-102-3003
NovaSeq 6000 S2 Reagent Kit (100 cycles)	Illumina	Cat#: 20012862
Chromium Single Cell 3' GEM, Library & Gel Bead Kit v3	10x Genomics	Cat#: PN-1000075
Chromium Single Cell B Chip Kit	10x Genomics	Cat#: PN-1000074
Deposited data		
Raw and processed transcriptome and barcode data	This manuscript	GEO: GSE173958
Analyzed lineage data	This manuscript	Mendeley Data: <a href="https://doi.org/10.17632/t98pjcd7t6.1">https://doi.org/10.17632/t98pjcd7t6.1</a>
Experimental models: Cell lines		
PDAC 6419c5 cells	Li et al. 2018	N/A
macsGESTALT PDAC cells	This manuscript	N/A
293T-V7 cells	This manuscript	N/A
293T-V8 cells	This manuscript	N/A
Experimental models: Organisms/strains		
Mouse: NOD scid	Jackson Laboratory	Cat#: 001303
Oligonucleotides		
Primer pairs (see Table S11)	This manuscript, IDT	N/A
Recombinant DNA		
pUltra-U6-gRNAs1-5	This manuscript	N/A
PB-EF1 $\alpha$ -Puro-V8.2	This manuscript	N/A
pLJM1-EGFP-V7	This manuscript	N/A
pLJM1-EGFP-V8	This manuscript	N/A
pCFDg1-5	This manuscript	N/A
pBS31-GFP-V8crRNAs-U6-tracr-Ub-M2rtTA	This manuscript	N/A
pUltra-U6-crRNAs-U6-tracr	This manuscript	N/A
p5xU6_5sgRNA-Hsp70-Cas9GFP-pA	Raj et al. 2018	N/A
pBS31	Beard et al. 2006	N/A
pUltra	Addgene	Cat#: 24129
pLJM1-EGFP	Addgene	Cat#: 19319
Lenti-iCas9-neo	Addgene	Cat#: 22667
psPAX2	Addgene	Cat#: 12260
pMD2.G	Addgene	Cat#: 12259
Super PiggyBac Transposase	SBI	PB210PA-1
Software and algorithms		
R v4.0.2	R Core Team	<a href="https://www.r-project.org/">https://www.r-project.org/</a>

10x Cell Ranger v3	10x Genomics	RRID: SCR_017344; <a href="https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger">https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger</a>
Monocle 3	Junyue Cao et al. 2019	RRID: SCR_018685; <a href="https://cole-trapnell-lab.github.io/monocle3/">https://cole-trapnell-lab.github.io/monocle3/</a>
Seurat v3.1.4	Stuart et al. 2019	RRID: SCR_016341; <a href="http://www.satijalab.org/seurat/">www.satijalab.org/seurat/</a>
tidyverse v1.3.0	Wickham et al. 2019	RRID: SCR_019186; <a href="https://CRAN.R-project.org/package=tidyverse">https://CRAN.R-project.org/package=tidyverse</a>
igraph v1.2.6	<a href="https://igraph.org/">https://igraph.org/</a>	RRID: SCR_019225; <a href="https://cran.r-project.org/web/packages/igraph/">https://cran.r-project.org/web/packages/igraph/</a>
ggraph v2.0.5	<a href="https://ggraph.data-imaginist.com/index.html">https://ggraph.data-imaginist.com/index.html</a>	<a href="https://cran.r-project.org/web/packages/ggraph/index.html">https://cran.r-project.org/web/packages/ggraph/index.html</a>
HOMER v4.11.1	Heinz et al. 2010	RRID: SCR_010881; <a href="http://homer.ucsd.edu/">http://homer.ucsd.edu/</a>
singscore v1.8.0	Foroutan et al. 2018	<a href="https://www.bioconductor.org/packages/release/bioc/html/singscore.html">https://www.bioconductor.org/packages/release/bioc/html/singscore.html</a>
survival v3.2-7	N/A	<a href="https://cran.r-project.org/web/packages/survival/index.html">https://cran.r-project.org/web/packages/survival/index.html</a>
inferCNV	Trinity CTAT Project	<a href="https://github.com/broadinstitute/inferCNV">https://github.com/broadinstitute/inferCNV</a>
Barcode alignment	McKenna et al. 2016	<a href="https://github.com/mckennalab/SingleCellLineage/">https://github.com/mckennalab/SingleCellLineage/</a>
TreeUtils	McKenna et al. 2016	<a href="https://github.com/mckennalab/TreeUtils">https://github.com/mckennalab/TreeUtils</a>

Lineage processing and analysis	This manuscript	<a href="https://github.com/ksimeono/macsgESTALT">https://github.com/ksimeono/macsgESTALT</a> & <a href="https://doi.org/10.17632/t98pjcd7t6.1">https://doi.org/10.17632/t98pjcd7t6.1</a>
Other		
Online tree browser	This manuscript	<a href="https://macsgestalt.mckennalab.org/">https://macsgestalt.mckennalab.org/</a>

## ***Resource availability***

### Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Christopher J. Lengner ([lengner@vet.upenn.edu](mailto:lengner@vet.upenn.edu)).

### Materials availability

Materials and reagents used in this study are listed in the Key Resources Table.

Reagents generated in our laboratory are available upon request. The plasmids needed to implement macsGESTALT will be made available through Addgene.

### Data and code availability

Raw and processed single cell lineage and transcriptional data are available through GEO: GSE173958. Further processed lineage data files and corresponding analysis scripts and R Notebooks are available together through Mendeley Data in a coherent file structure: <http://dx.doi.org/10.17632/t98pjcd7t6.1>. R Notebooks and scripts alone are also available through Github: <https://github.com/ksimeono/macsgestalt>.

## ***Experimental models and subject details***

### Cell lines

All cells were cultured in a 5% CO<sub>2</sub> incubator at 37 °C in culture media (High Glucose DMEM, 10% FBS, 1% glutamine with penicillin and streptomycin). 293T cells were a gift from Dr. Jeremy Wang at the University of Pennsylvania. Barcoded 293T cells for the gRNA screen were produced by infecting with pLJM1-EGFP-V7 or pLJM1-EGFP-V8 lentivirus at low MOI (MOI < 0.2) and sorted by fluorescence-activated cell sorting (FACS) for GFP using a BD FACSAria II (BD Biosciences).

For the PDAC cells used to generate macsGESTALT PDAC cells, we selected the most metastatically aggressive cell line (6419c5) from a published library of clonal PDAC lines ([J. Li et al. 2018](#)), which were each derived from harvested KPCY tumors. While this cell line originated from a single cell bottleneck during derivation, it had since been passaged ~15x, thereby overtime in culture, becoming effectively polyclonal at the point of macsGESTALT barcode delivery.

macsGESTALT components were introduced into PDAC cells in 3 steps: First, dox-inducible Cas9 was integrated with Lenti-iCas9-neo (Addgene #22667) ([Jian Cao et al. 2016](#)), and infected cells were selected for neomycin resistance via G418 for 7 d. Second, the cells were infected with pUltra-U6-gRNAs1-5 at high MOI (MOI > 0.8), and the top 50% of GFP positive cells were sorted by FACS using a BD FACSAria II. This step was repeated once to produce cells with high gRNA array expression to ensure a high editing rate. This can be decreased to slow and spread the editing rate over time.

Third, cells from the previous steps were barcoded by cotransfecting PB-EF1 $\alpha$ -Puro-V8.2 library and Super PiggyBac Transposase plasmid (SBI #PB210PA-1) at a 1:10 molar ratio using Lipofectamine 3000 (ThermoFisher). Barcoded cells were puromycin-selected for 7 d. To maintain diversity and limit leaky editing, cells were expanded after withdrawal of puromycin and frozen down with minimal time in culture (< 7 d). For lineage tracing experiments, cells were only expanded after thawing for 2-4 d as needed prior to orthotopic injection or experiment start.

### Mice

NOD scid male mice were acquired from Jackson Laboratory. 10 week old mice were used for orthotopic injection. All mice were maintained in a specific pathogen-free environment at the University of Pennsylvania Animal Care Facilities. All experimental protocols were approved by and performed in accordance with the relevant guidelines and regulations of the Institutional Animal Care and Use Committee of the University of Pennsylvania.

## ***Method details***

### Plasmid design and construction

All Gibson assemblies were performed using NEBuilder HiFi DNA Assembly Master Mix (NEB #E2621) and were assembled at 50 °C for 60 min at appropriate molar ratios. For cloning, all PCRs were performed using HotStart ReadyMix (Kapa Biosystems #KK2601). Restriction enzymes, instead of PCR, were used to linearize vector backbones to prevent backbone mutations. All bacterial transformations were performed with NEB Stable Competent E. coli (NEB #3040H) and cells were grown at 30 °C for 24 h, unless otherwise noted. Final plasmid preps were performed with Zymopure II Plasmid Kits (Zymo Research #D4202). All regulatory, coding, and editing-related regions in final assembly products were validated by Sanger sequencing. All gene block sequences were ordered from IDT.

V7 and V8 barcoding lentiviral transfer plasmids used for guide RNA array screening were constructed in 2-part Gibson assemblies using pLJM1-EGFP (Addgene #19319) ([Sancak et al. 2008](#)) backbone digested with EcoRI + gene blocks for V7 or V8 barcodes to make pLJM1-EGFP-V7 and pLJM1-EGFP-V8.

pUltra-U6-crRNAs-U6-tracr was constructed in a 3-part Gibson assembly using PacI linearized pUltra (Addgene #24129) ([Lou et al. 2012](#)) backbone, a U6-driven array of 10 V8 targeting crRNAs (crRNAs) interspersed by tRNAs ordered as a gene block (pUltra5-U6crRNA-GA1), and another gene block encoding a U6-driven tracrRNA (GA1-U6-tracr-pUltra3).



The dox-inducible crRNA array plasmid, pBS31-GFP-V8crRNAs-U6-tracr-Ub-M2rtTA, was constructed in a 3-part Gibson assembly using EcoRI linearized pBS31 ([Beard et al. 2006](#)), a gene block containing 10 V8 targeting crRNAs interspersed by tRNAs in the 3' of a GFP opening reading frame (ORF) (TP-gB-1), and a gene block containing U6-driven tracrRNA followed by Ubc promoter-driven M2-rtTA with a V8 barcode of 10 targets in the 3' UTR (TP-gB-2). The barcode was excised for transient transfection gRNA screening experiments by digesting with Nsil and religating the backbone.

p5xU6\_5sgRNA-Hsp70-Cas9GFP-pA that had V7 gRNAs 5-9 each with a separate U6 promoter was a gift from J. Gagnon ([Raj et al. 2018](#)).

pCFDg1-5 gRNA-tRNA array was constructed stepwise as previously described using pCFD5 (Addgene #73914) ([Port and Bullock 2016](#)) as a template and V8 targeting gRNAs.

pUltra-U6-gRNAs1-5 lentiviral transfer plasmid, which was used to make macsGESTALT PDAC cells, was generated in a 3-part Gibson assembly using pUltra backbone linearized with PacI, a gene block with U6 promoter and gRNA 1 (pUltra5-U6-gRNA1), and a PCR-amplicon, amplified from pCFDg1-5, containing gRNA-tRNAs 2-5 (gRNAs1-5-pUltra3), thereby producing a constitutively-expressed five gRNA-tRNA array and a constitutive GFP selection marker.

PB-EF1 $\alpha$ -Puro-V8.2 library cloning was performed as a 3-part Gibson assembly: 1) PB-CMV-MCS-EF1 $\alpha$ -Puro (Systems Biosciences PB-510B-1) was digested with SpeI and HpaI to excise its cargo and create a linear backbone. 2) EF1 $\alpha$  promoter and puro resistance gene were amplified from lentiGuide-Puro (Addgene #52963). 3) The V8.2 target array was ordered as a gene block. This assembly produced the PB-EF1 $\alpha$ -Puro-V8.2 vector. Then, the barcode library was generated via a 2-part Gibson assembly using EcoRI linearized PB-EF1 $\alpha$ -Puro-V8.2 and a random 10 bp containing staticID (static barcode) fragment, which was made by annealing and extending a pair of oligos (targetbarcode-r:

TTTGTCCAATTATGCTCGAGGTCGAGAATTNNNNNNNNNNCGTTGATCGCACGCCA, targetbarcode-f2: TAGTTGGTTCCTACTGGCGTGCGATCAACG). The library was transformed into NEB 10-beta Electrocompetent E. coli (NEB #3020K), and the entire transformation was grown as a midi culture and prepped with Chargeswitch Pro Filter Midi Kit (Thermofisher #CS31104).

#### Viral production

Lentiviruses were packaged in HEK 293T cells using psPAX2 (Addgene #12260) and pMD2.G (Addgene #12259) second generation packaging and envelope plasmids. Viral supernatants were collected 2-4 d post-transfection and filtered through 0.45  $\mu$ m filters. Filtered supernatants were either stored at -80  $^{\circ}$ C (never refrozen) or used fresh to infect cells.

#### Guide RNA array editing screen

293T cells barcoded with pLJM1-EGFP-V7 or pLJM1-EGFP-V8 lentivirus were transiently transfected with different combinations of plasmids to test gRNA array editing efficacy. Barcoded cells plated at 250,000 cells per well of 6-well plates, and transfected the following day with Lipofectamine 2000 (Thermofisher #11668030). 1.5 µg of px330 was used in each well (except no-transfection and pUltra-only control wells). All wells receiving a gRNA array plasmid were also transfected with a 1:1 molar amount of the appropriate gRNA plasmid compared to px330. Dox was initiated where appropriate the day after transfection. Additionally, as a positive control, one well received px330 and *in vitro* transcribed (IVT) gRNAs. Guide templates matching the V8 target sites were constructed and transcribed using GeneArt Precision gRNA Synthesis Kit (Thermofisher #A29377); gRNA 6 and 7 IVT reactions failed and these guides were excluded from further steps. IVT gRNAs were transfected using Lipofectamine CRISPRMax (Thermofisher #CMAX00001) 24 h after px330 was transfected. Expression of plasmids containing fluorescent markers was confirmed by microscopy. Cells were then allowed to expand and edit for one week and then harvested for library preparation and sequencing.

#### PDAC dox-induced *in vitro* editing experiments

PDAC cells were cultured in complete media (DMEM, 10% FBS, 1% glutamine with penicillin and streptomycin). Dox-induced editing checks of macsGESTALT PDAC cells were performed in two separate experiments: In the first experiment, cells were plated and started on dox at 3 doses, 0, 0.1, or 2 µg/mL, with media change every other day. Cells were collected at 2 timepoints — after 1 and 2 weeks of dox exposure — and harvested for library preparation and sequencing. In the second experiment, cells were

kept on 6 different dosages of dox, 0, 10, 50, 100, 500, or 1,000 ng/mL, for 2 weeks and harvested for library preparation and sequencing. Prior to the start of editing experiments, cells experienced 3 weeks of culture time during barcode drug selection, expansion, and freeze/thawing, during which time background editing from leakiness was possible.

#### Bulk DNA barcode sequencing

For all bulk DNA editing experiments, approximately one million cells were harvested per condition, washed, pelleted, and genomic DNA extracted with the NucleoSpin DNA RapidLyse Kit (Macherey-Nagel #740100.50). Genomic DNA was normalized to 30-50 ng/ $\mu$ L for each sample. All PCR reactions were performed using SYBR-containing master mix from the KAPA Real-Time Library Amplification Kit (Kapa Biosystems #KK2702) and terminated in the mid-exponential phase to limit over-amplification. AMPure beads (Agencourt Beads, Beckman Coulter #A63880) were used at a ratio of 1.5x to purify products after all PCR reactions. Barcodes were amplified from genomic DNA in a nested approach and sequencing adaptors, sample indices, and flow cell adaptors were added by a series of subsequent PCRs. For 293T samples containing pLJM1-EGFP-V7 or pLJM1-EGFP-V8, barcodes were amplified and adaptors added in a series of 3 PCRs. For PDAC samples containing PB-EF1 $\alpha$ -Puro-V8.2, barcodes were amplified and adaptors added in a series of 4 PCRs. Primer sequence, purpose, and annealing temperature for all PCRs in both of these library preparations are included in **Table S6**. In all cases, 250 ng of genomic DNA was loaded into a 50  $\mu$ L PCR. Sample indices were added using NEBNext Multiplex Oligos for Illumina (Dual Index Primers Set – New England Biolabs). The concentration of final amplicons was measured by Qubit

and the length validated by TapeStation HSD1000 prior to sequencing using Illumina MiSeq 600-cycle v3 Reagent Kits with the following run parameters: Read 1 - 301 cycles, i7 index - 8 cycles, i5 index - 8 cycles, Read 2 - 301 cycles. Bulk sequencing data for all samples was aligned and processed as previously reported ([McKenna et al. 2016](#)) and available as a docker image <https://github.com/mckennalab/SingleCellLineage/>, with the UMI option set to FALSE (no UMI used). Output files were used for generating visualizations using the R programming language.

#### Limiting dilution PDAC experiments

macsGESTALT PDAC cells were plated in a limiting dilution of approximately ~5 or ~100 cells per well in a 48-well plate. Single cells gave rise to colonies and expanded. Cells were all allowed to expand without split for 2 weeks. The 100-cell wells were confluent and overgrown after 1 week in culture. The 5-cell wells were approximately 80-90% confluent at 2 weeks. At 2 weeks, a healthy, representative well from each condition was selected and passaged at a 1:2 split into a well of a 6-well plate. After 3 d, cells were harvested and dissociated using 500  $\mu$ L TrypLE (Thermofisher #12605010) for 3-5 min. Reactions were neutralized with 3 mL culture media. Cell clumps were further dissociated by gently pipetting up and down 10x with a p1000, and then cells were centrifuged at 250g for 5 min. Cells were gently resuspended with a p1000 in 1 mL culture media, filtered through a 30  $\mu$ m strainer, ensured to be in a single cell suspension under a light microscope, and counted with a hemocytometer. Cells were washed twice with 1 mL cold HBSS with 0.04% BSA (centrifuged at 150g for 3 min each time). Cells were filtered again through a 30  $\mu$ m strainer and resuspended in cold HBSS

with 0.04% BSA at a concentration of 700 cells/ $\mu$ L. Cells were counted again with a hemocytometer to ensure accurate concentration. For the 5-cell dilution sample, 8,000 cells were loaded on 10x (Chromium Single Cell 3' Reagent Kits v3) targeting 5,000 cell recovery; for the 100-cell dilution sample, 16,000 cells were loaded targeting 10,000 cell recovery.

#### Orthotopic metastasis model

macsGESTALT PDAC cells were thawed and expanded for 2-4 d prior to dissociation and orthotopic injection into 10 week old NOD scid male mice. Approximately 30,000 PDAC cells were injected into the surgically-exposed tail of the pancreas, as previously described in detail ([Aiello, Rhim, and Stanger 2016](#)). Cells were allowed to engraft; then doxycycline was initiated 1 week post-injection and given continuously in the drinking water at 1 mg/mL. Mice were harvested at approximately 5 weeks post injection, once reaching morbidity. Primary tumor (PT), liver, lung, peritoneal macrometastases, and surgical-site lesions were sorted for both mice. Due to a more productive blood-draw, circulating tumor cells (CTCs) were captured for M1 but not M2. Additionally, the surgical-site lesion, which is similar in size and location to other peritoneal macrometastases, was processed separately in M1 but not M2..

#### Blood harvest and preparation

When harvesting tissues, blood was extracted first via cardiac puncture using a 25 gauge 5/8 needle with 1 mL syringe attached. A successful blood draw was 400-700  $\mu$ L, which was immediately transferred to a FACS tube containing 4% sodium-citrate in Milli-Q water. This was pelleted at 500 g for 5 min and red blood cells were lysed by

resuspension in 2 mL ACK (Ammonium-Chloride-Potassium) buffer and incubation for 5 min at room temperature. 3 mL PBS were added and the mix was pelleted at 500 g for 5 min. Red blood cell lysis was repeated 2 times. Finally, cells were resuspended in 400  $\mu$ L of cold FACS buffer (PBS, 2% FBS, 1 mM EDTA, 40  $\mu$ g/mL DNase) with DAPI and strained through a 35  $\mu$ m filter for FACS.

#### Macro lesion harvest and dissociation

Primary tumor and macrometastases (metastases that could be manually handled, including surgical-site lesion) were excised from surrounding tissue, removing as much normal surrounding tissue as possible. All macrometastases from a mouse were processed as one sample. Samples were then transferred to a 6-well plate and washed with cold PBS 3x. Samples were minced, then transferred into 10 mL of DMEM containing 2 mg/mL collagenase IV plus 40  $\mu$ g/mL DNase and incubated in a 37 °C shaker for 30 min. Cells were isolated by physical dissociation, filtered through a 70  $\mu$ m cell strainer, and neutralized with cold DMEM. Samples were centrifuged at 350g for 5 min and resuspended in 500  $\mu$ L cold FACS buffer (above). Cells were centrifuged at 350g for 5 min, resuspended in 1 mL cold FACS buffer with DAPI, pipetted up and down 5x gently with p1000, and strained through a 35  $\mu$ m filter for FACS. Samples and cells were kept on ice throughout unless otherwise indicated.

#### Liver and lung harvest and dissociation

To minimize blood contamination in the liver and lungs, 25 mL of cold PBS was perfused into the right ventricle of the heart (after blood draw from the heart). The entire liver (any macrometastases near the liver surface were completely excluded) and lungs were

excised and processed identically to PTs, until immediately following the 30 min shaking digestion step. Here, samples were filtered through 100  $\mu$ m cell strainers and then neutralized and centrifuged as with PTs, except 250g was used instead of 350g for centrifugation steps.

Liver samples were resuspended and further digested in 5 mL TrypLE for 5 min at 37 °C. Digestions were neutralized with cold DMEM + 10% FBS, centrifuged at 250g for 5 min, resuspended in 3 mL ACK, and incubated for 3 min at RT. Liver reactions were neutralized with cold PBS, centrifuged at 250g for 5 min, resuspended in 5 mL cold FACS buffer with DAPI, pipetted up and down 5 times gently with p1000, and strained through a 35  $\mu$ m filter for FACS.

Lung samples were processed identically to liver samples except the order of ACK and TrypLE digestion steps was reversed (ACK before TrypLE). Additionally, lung samples were much smaller than liver samples and were thus only resuspended in 500  $\mu$ L of cold FACS buffer with DAPI for FACS. Both liver and lung samples were kept on ice throughout unless otherwise indicated.

#### Cancer FACS sorting and 10x Chromium loading

Cancer cells were isolated from dissociated tissues via FACS using a BD FACSAria II. After gating for singlets and live cells, GFP+ cells were sorted, thereby purifying PDAC cells from normal cells. For samples with a high yield of cells (PT, macrometastases, surgical-site), 30-35,000 cells were sorted on the purity setting. For each of the lung, liver, and blood samples, the entire sample was sorted on the yield setting to recover as



many GFP+ cells as possible. The liver for M1 was stopped with 20% of the sample volume remaining due to excessively long sorting time. Cell numbers recovered for lung and liver were similar for each mouse (M1 liver: 22,000 (80% of total), M2 liver: 30,000, M1 lung: 1,000, M2 lung: 1,500).

After sorting, all samples were passed through a 30  $\mu$ m filter and then centrifuged at 500g for 5 min and checked for visible pellets. Supernatant was removed to leave 20-30  $\mu$ L of solution to not disturb the pellets. Remaining volume was measured and raised to 50  $\mu$ L total by adding a 1:1 mixture of cold FACS buffer (without DNase) and nuclease-free water. 46.6  $\mu$ L of these samples was loaded for 10x (Chromium Single Cell 3' Reagent Kits v3), thereby superloading some lanes with up to 25-30,000 cells (macsGESTALT single cell barcode sequencing allows explicit detection of multiplets, see **Figure S2J** and Methods subsection "Clonal reconstruction and multiplet elimination").

#### Single cell transcriptome sequencing

Single cell RNA-seq libraries were prepared as in the 10x Chromium Single Cell 3' v3 user guide (Rev A) until Step 2.3. After cDNA amplification, the 100  $\mu$ L cDNA PCR was split 50:50 for separate barcode and transcriptome library preparation. Transcriptome library construction continued as in the 10x user guide instructions. Indexed and pooled single cell transcriptome libraries for each mouse were sequenced separately on the NovaSeq 6000 System with S2 100-cycle kits.

#### Single cell barcode sequencing

For all single cell barcode PCRs (as for bulk DNA barcode PCRs), SYBR-containing master mix from the KAPA Real-Time Library Amplification Kit was used, and PCRs were stopped in mid-exponential phase. All primers were used at 10  $\mu$ M. Primer sequence, purpose, and annealing temperature for all library preparation PCRs are included in **Table S6**.

The barcode split of the cDNA amplification reaction (from 10x Single Cell 3' v3 Step 2.2) was purified via 1.2x SPRI Select (Beckman Coulter #B23317). cDNA products were eluted in 40  $\mu$ L of EB. Concentrations were measured by Qubit, and 2 ng/ $\mu$ L dilutions in EB were created for each sample. Barcode amplification and adaptor and sample index addition were performed in 2 sequential PCRs.

Barcodes were selectively amplified by PCR1. Here, 50 ng of each purified, diluted cDNA amplification sample was used to template a 100  $\mu$ L PCR. After mixing, the reaction was split into 4 smaller reactions of 25  $\mu$ L each for cycling. PCR cycling conditions were 1) 95  $^{\circ}$ C for 3 min, 2) 14-15 cycles of 98  $^{\circ}$ C for 20 s, 65  $^{\circ}$ C for 15 s, 72  $^{\circ}$ C for 15 s. Sample reaction splits were re-pooled after cycling, and products were purified with 0.9x SPRI Select and eluted in 60  $\mu$ L EB.

Sample indices were added in PCR2. Here, 5-10  $\mu$ L of the eluted products of PCR1 (1:12 or 1:6 overall dilution) were used to template a 100  $\mu$ L PCR, which was again mixed and split into four smaller reactions of 25  $\mu$ L each. PCR cycling conditions were 1) 95  $^{\circ}$ C for 3 min, 2) 6 cycles of 98  $^{\circ}$ C for 20 s, 65  $^{\circ}$ C for 15 s, 72  $^{\circ}$ C for 15 s. Sample reaction splits were re-pooled after cycling. Dual-sided size selection of complete

barcode amplicons was performed using SPRI Select at an exclusion ratio of 0.5x and a selection ratio of 0.7x. Amplicons were eluted in 32  $\mu$ L EB.

Barcode library size and concentration were checked via TapeStation HSD5000 and Qubit, respectively. Libraries were sequenced using Illumina MiSeq 600-cycle v3 Reagent Kits with the following run parameters: Read 1 - 28 cycles, i7 index - 8 cycles, Read 2 - 500 cycles. M1 was sequenced with 3 kits. Since barcode recovery only increased 5-10% with two additional kits for M1, M2 barcode library was sequenced with a single kit. Limiting dilution experiment libraries were also sequenced with a single kit.

## ***Quantification and statistical analysis***

### Single cell transcriptome data processing

Single cell transcriptome sequencing data was aligned and processed using 10x Cell Ranger v3.1 with the mm10 reference genome. Filtered matrices from Cell Ranger output were further processed using Seurat 3.1.4 (<https://satijalab.org/seurat/>) (Stuart et al. 2019). All samples across both mice were merged into a single Seurat object. Low quality cells with  $\leq 1,000$  genes or  $\geq 0.20$  mitochondrial gene fraction (mito fraction) were filtered out. Cell cycle score and phase were determined for each cell using the CellCycleScoring function ([https://satijalab.org/seurat/v3.1/cell\\_cycle\\_vignette.html](https://satijalab.org/seurat/v3.1/cell_cycle_vignette.html)).

Variable feature selection, scaling, and normalization were performed using SCTransform, while regressing cycle scores and mito fraction. Dimensionality reduction by PCA was performed using the first 15 principal components (PCs). Cells were plotted in UMAP space and a clearly-separated, large cancer cell cluster was observed, distinct from smaller clusters of contaminating normal cells, mostly derived from samples sorted on the FACS yield setting. Contaminating normal cells were filtered out. 10x cell barcodes, here referred to as cellIDs, for the cancer cells were then exported and used for initial macsGESTALT barcode data filtering.

### Single cell lineage data processing

Single cell barcode sequencing data was aligned, collapsed by UMI, and processed, as previously reported (McKenna et al. 2016) via a pipeline available as a docker image here: <https://github.com/mckennalab/SingleCellLineage/> and described further here:

<https://github.com/ksimeono/macsgESTALT>. For each sample, stats files, containing aligned and collapsed edited barcode sequence data, were extracted from pipeline output and used for clonal and subclonal analysis in R v4.0.2 and tidyverse v1.3.0 ([Wickham et al. 2019](#)). Sample stats file for different harvest sites from a mouse were merged. However, each mouse and limiting dilution experiment was processed separately.

To ensure high-quality barcode data was used for reconstruction, five initial filtering steps were applied: First, cellIDs not present in the initial transcriptome cellID list (or v3 10x whitelist for limiting dilution experiments without transcriptional data) were filtered. Second, transcripts (UMIs) with incomplete static barcode (staticID) sequences were filtered. Third, staticIDs with less than two UMIs per cell were removed. Fourth, staticIDs with less than two UMIs per cell on average were filtered. Fifth, staticIDs found in less than 5 cells were filtered. Specific thresholds were determined by examining elbow plots of the relevant parameters (see <https://github.com/ksimeono/macsgESTALT> for detailed R Notebooks with inline plots for each mouse).

#### Clonal reconstruction and multiplet elimination

Next, potential clonal groupings of cells based on staticID content (absence or presence) were identified by complete-linkage hierarchical clustering. The staticID content of resulting clusters was examined, and clusters were found to be often improperly fractured due to cells with undetected staticIDs. To identify real clones defined by sets of staticIDs, clustering results were pruned by excluding clusters of less than five cells and staticIDs found in less than 20% of cells for a particular cluster (see

<https://github.com/ksimeono/macsgESTALT> for relevant visualizations and code). For clusters of less than 20 cells, staticIDs found in less than 35% of cells were further excluded. Then, clusters that were either duplicates or subsets of other clusters in terms of their defining staticIDs were collapsed. Finally, remaining staticID cluster sets were manually inspected for improperly fractured clusters, and any remaining improper cluster splits were merged or collapsed (usually this was either not necessary or was only needed for a few clusters).

After cluster cleanup, staticID sets were extracted and used to assign cells. Cells were matched to clusters based on their staticIDs. This process also served to explicitly identify interclonal multiplets, i.e. if a cell matched two or more clusters, this cell was removed as a multiplet. This method performed well, as only a small fraction of cells, ranging from 0 to 0.54% across experiments, went unmatched. Unmatched cells likely belonged to very small clones, only found in *in vivo* experiments. Furthermore, the percentage between mice was strikingly consistent (M1: 0.54% and M2: 0.51%), highlighting the reproducibility of the cancer model system and reconstruction approach. Only matched singlets were retained for downstream analysis.

With this orthotopic model, it is possible that some of the cells injected can leak out of the pancreas during and after injection and directly colonize the peritoneal cavity (although we sought to minimize this as previously described ([Aiello, Rhim, and Stanger 2016](#))). To eliminate any such cells from further analysis, we filtered clones that were detected in disseminated sites but not in the PT. This resulted in the removal of a small

number of cells (M1: 1.49% and M2: 0%) from a few clones only found in peritoneal macrometases and in the surgical site lesion of M1.

In a true singlet, without genomic duplication of a barcode, each cellID-staticID pair should have a single mutagenized allele. To detect potential intraclonal multiplets or duplicated barcodes, we calculated the number of unique mutagenized evolving barcodes for a cellID-staticID pair, and mutagenized barcodes with less than 25% of the UMIs for that cellID-staticID pair were removed as technical noise.

PDAC is known to undergo large-scale copy-number changes via chromosomal instability. We observed this in our CNV analysis using InferCNV (**Figure S3B**). While most staticIDs had a median of one mutated allele per cell, some had a median of two and a notably higher average. We speculated that these might be barcodes that resided in genomic areas that underwent copy number gain at some point after barcode integration. StaticID that had an average of 1.3 or greater mutated alleles per cell were considered to be potentially duplicated or triplicated.

Per 10x Chromium 3' Single Cell v3 documentation (page 16), our overall expected multiplet rate for *in vivo* experiments with superloading was approximately 12% to 15%. Having explicitly detected and filtered interclonal multiplets, we next removed potential intraclonal multiplets. We filtered all cells with an average number of unique mutated alleles per staticID greater than 1.25, except for cells containing a potentially duplicated staticID; for these cells, the threshold was less stringent, at greater than 3. This resulted

in appropriate overall multiplet rates of 12% for M1 and 15.7% for M2. Only true singlets were retained for further analysis.

After these filtering steps, clones that were detected in disseminated sites but not in the PT were again removed if present, and clones were then numbered by their size in the primary tumor, largest to smallest. These rankings are used to refer to clones throughout the paper with the mouse number appended, i.e. M1.1 or M2.14. These finalized clones were used for calculating clone size and clone fraction for each harvest site. These final filtered, clone-assigned singlets were used for further single cell transcriptional analysis.

Clonal aggression scores were estimated by giving points for size and fraction. For each non-PT harvest site where a clone was present 0.5 points were awarded. If the clone's fraction was higher at a disseminated site than at the PT than it was rewarded an additional 1 point for that site. If a clone made up 5% or more of a disseminated site it received an additional 0.5 points for that site and a further 0.5 points if it was 10% or more.

For limiting dilution validation experiments, cells were visualized by their static barcode expression using tSNE in Seurat. A static barcode (rows) by cells (columns) expression matrix was generated. Just as in a regular transcriptome scRNAseq analysis, this matrix was used to generate a SeuratObject, where static barcodes were treated as features. The first 50 dimensions were used for tSNE plotting.

#### Single cell transcriptional analysis



Transcriptional analysis continued using only singlets with quality barcode information (from above section). Seurat objects were converted into cell\_data\_set objects, and Monocle 3 (<https://cole-trapnell-lab.github.io/monocle3>) was used for all further transcriptional analysis. Preprocess\_cds was run with top 20 dimensions (PCA) and align\_cds was run with batch correction for harvest site and regression for cycle scores and mito fraction. Cells were plotted in UMAP space and two clusters of low quality or contaminating cells were removed. The first was a cluster of cells distinguished by high ribosomal fraction that was derived from cells of many clones and harvest sites. These cells were likely technical artifacts observed from droplet library preparation. The second was a cluster of cells with high hepatic gene expression. These cells derived from primarily the liver harvest sites and were most likely contaminating tumor-liver multiplets that had escaped initial filtrations steps.

Following these filtrations, preprocess\_cds and align\_cds were run again as before but with the top 25 dimensions, as determined by examining an elbow plot using plot\_pc\_variance\_explained. Cells were plotted in UMAP space and clusters found using cluster\_cells. Further transcriptional analyses and visualizations on all mouse cancer cells together were performed using Monocle 3 functions and custom R scripts as needed. For analyses on individual mice, cells were extracted and reprocessed as above but with the top 20 dimensions by PCA.

#### Copy-number variation (CNV) analysis

InferCNV was used for single cell CNV analysis

(<https://github.com/broadinstitute/inferCNV/wiki>). Default settings were used. Cutoff = 0.1

was used, which is recommended by InferCNV for 10x data. Clones were treated as cell groups, with `cluster_by_groups = T`. Clones with >200 cells were downsampled to 200. For clones  $\leq 200$  cells, all cells were included.

### PseudoEMT analysis

PseudoEMT or pseudotime analysis was performed by finding a trajectory in UMAP space using `learn_graph` with default settings. The root (most epithelial region) was placed where epithelial gene expression peaked. This additionally led to the most mesenchymal region existing at the end of the trajectory, thus resulting in a pseudoEMT spectrum. To find genes whose expression varied significantly along pseudoEMT, `graph_test` was used with the 'principal\_graph' parameter selected. The top 3000 genes were retained, all of which had  $q \sim 0$  and Moran's  $I > 0.1$  (**Table S2**). For the top 3000 genes, kinetic expression curves were clustered into groups by ward.D2 clustering using the R Pheatmap package, and the resulting tree was cut into six groups, which were named in order from epithelial to hybrid to mesenchymal patterns of expression.

To find enriched transcription factor motifs within the six gene clusters, `findMotifs.pl` from HOMER was used with the provided mouse promoter set. All default parameters were used, except for promoter region (-500, 50 bp from TSS) and background promoter frequency (derived from all top 3000 pseudoEMT genes). Known motifs passing an enrichment cutoff of  $p < 0.05$  were extracted. The target genes of each motif were obtained using HOMER's `annotatePeaks.pl`. Also for each pseudoEMT gene group, molecular signature database (mSigDB) gene set enrichment was determined using the hypergeometric test within HOMER.

### Subclonal and phylogenetic reconstruction

Using filtered barcode data (from material and Methods subsection "Clonal reconstruction and multiplet elimination"), duplicated barcodes were removed entirely (this also removed any cells whose only recovered barcodes were duplicated). Cells with greater than one unique mutated allele per staticID were then filtered. For each cell in a clone, a barcode-of-barcodes was generated by concatenating all evolving barcode alleles, ordered by staticID. If a cell was missing a staticID, 'UNKNOWN\_UNKNOWN\_UNKNOWN\_UNKNOWN\_UNKNOWN' was concatenated for that staticID to note the missing information for all five target sites. Thereby, for an example clone defined by four staticIDs, every cell had four evolving barcodes concatenated in order and 20 target sites overall, including any missing information.

Within each clone, cells with identical barcode-of-barcodes were then grouped into subclones of indistinguishably closely related cells. To limit computational time required for downstream phylogenetic reconstruction of subclonal relationships, we pruned subclones of only a single cell from the largest clones, i.e. clones with  $\geq 50$  cells. This greatly increased computational efficiency while still retaining meaningful subclones.

Separate files were constructed for each clone, containing subclones with associated barcode-of-barcodes alleles. Phylogenetic reconstruction of subclonal relationships was performed for each clone barcode-of-barcodes file separately via TreeUtils (<https://github.com/mckennalab/TreeUtils>). TreeUtils performs reconstruction using

Camin-Sokal maximum parsimony via the PHYLIP Mix software package ([Felsenstein 1989](#)), as previously described in depth ([McKenna et al. 2016](#)).

Further analysis then resumed in R. Clone Newick files were extracted from TreeUtils output and converted to an edgelist dataframe format. Clone edgelists were combined into a single large edgelist with a common root node (for each mouse separately). A small fraction of clones that were entirely defined by staticIDs that had been genomically duplicated, and were thus left out of phylogenetic analysis, were added back as a single node emerging directly from the root. At this point, cellIDs were added as terminal nodes emerging from subclone nodes (or directly to clone nodes for clones that were left out of phylogenetic analysis due to barcode copy gain). Cell nodes were then annotated with harvest site, transcriptional, and other information as needed. For circle pack or tree visualization, edgelist dataframes were converted to igraph graph objects (<https://igraph.org/r/>) and plotted using ggraph (<https://github.com/thomasp85/ggraph>).

#### Subclonal dissemination calculation

Shannon's Equitability ( $E_H$ ) was used as a statistical measure of dissemination across harvest sites. To calculate  $E_H$ , Shannon Diversity ( $H$ ) was first calculated as follows:

$$H = - \sum_{i=1}^S p_i * \ln(p_i)$$

$S$  is the number of distinct harvest sites analyzed (six for M1, four for M2).  $p$  is the sampling normalized proportion at which a subclone is recovered from a harvest site, i.e. if a subclone is only found in the PT,  $p_{PT} = 1$ , while  $p = 0$  for all other sites. A subclone's  $H$  is then used to calculate its  $E_H$  as follows:

$$E_H = \frac{H}{H_{max}} = \frac{H}{\ln(S)}$$

$E_H$  therefore normalizes  $H$  by the number of harvest sites analyzed to exist between 0 and 1, with 1 being completely even dissemination and 0 being no dissemination. For example, a subclone found at only one harvest site is not metastatically aggressive and has an  $E_H = 0$ .

#### PseudoEMT across ancestral relationships

Comparison of pseudoEMT for root clades, subclones, and cells was performed in R. To determine root clade pseudoEMT values, we recursively calculated the weighted mean pseudoEMT value of ancestral nodes moving backwards along phylogenetic trees. Root clades were the nodes immediately preceding the common root of M1.1. These clades are depicted by the outermost circles in the circle packing visualizations of M1.1 (**Figures 5A and 5B**). The density of root nodes, subclones, and cells along the pseudoEMT axis was then plotted as a ridge plot for comparison.

#### Identifying genes associated with dissemination

Regression of  $E_H$  against single cell gene expression was performed while regressing out harvest site, cell cycle scores, and mito fraction. Genes with  $q < 0.05$  and greater than 1000 total transcripts across all cells were retained for further analysis. For analysis of highly expressed and highly associated genes, only genes with greater than 50,000 total transcripts and an absolute estimate of association greater than 0.1 were retained.

#### TCGA survival analysis

PseudoEMT genes (n = 3000, M1) and genes associated with dissemination (n = 2010, M2) were mapped to their human homologs using getLDS() from the biomaRt package. All homologous genes were included. Preprocessed transcriptomic data (FPKM abundance after upper quantile normalization; FPKMuq) ) from TCGA (<https://www.cancer.gov/tcga>) for patients with pancreatic adenocarcinoma (TCGA-PAAD; n = 173), breast invasive carcinoma (BRCA; n=969), lung adenocarcinoma (LUAD; n=526), colon adenocarcinoma (COAD; n=517) or prostate adenocarcinoma (PRAD; n=541) were obtained using the R package TCGAbiolinks.

Using the singscore package (Foroutan et al. 2018), patients' enrichment scores were determined for either each pseudoEMT gene cluster (E, H1, H2, H3, H4, M) or genes positively vs negatively associated with aggression. Patient survival (from the time of pathological diagnosis) was obtained from TCGA clinical data for each cancer.

Univariate and multivariate Cox regression analysis was performed in the R environment (survival) to determine the hazard associated with either the pseudoEMT gene signatures (M1) or dissemination (M2) for each cancer. Wald test, LLR and Score test were all significant ( $p < 0.05$ ), indicating the regression models were significant.

#### Pseudobulk and metagene analyses

The aggregate\_gene\_expression function from Monocle 3 was used to perform pseudobulk and metagene analyses. For testing whether clones retained their transcriptional identity, pseudobulk samples consisting of clone and harvest site combinations were generated, and only pseudobulk samples with >20 cells were used for further analysis. The entire transcriptome for each pseudobulk sample was

aggregated and used to hierarchically cluster samples via the Pheatmap package, with the ward.D2 clustering option.

### **Additional resources**

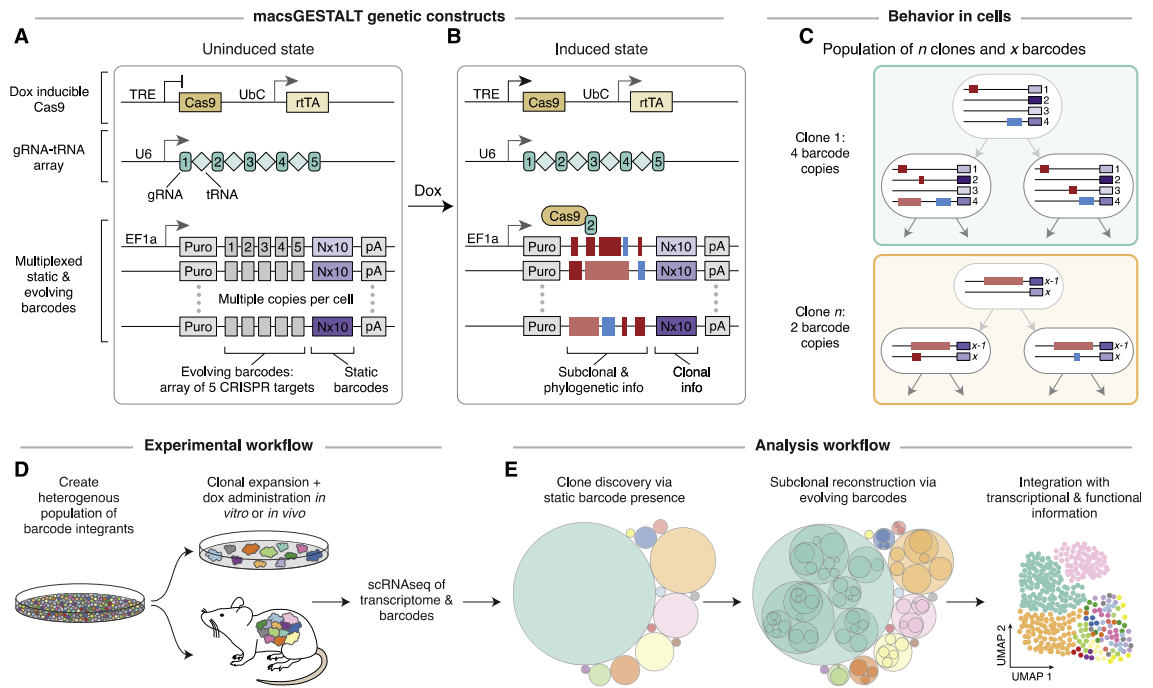
Interactive online browser of the lineage relationships reconstructed in this study:

<https://macsgestalt.mckennalab.org/>.



## Figures

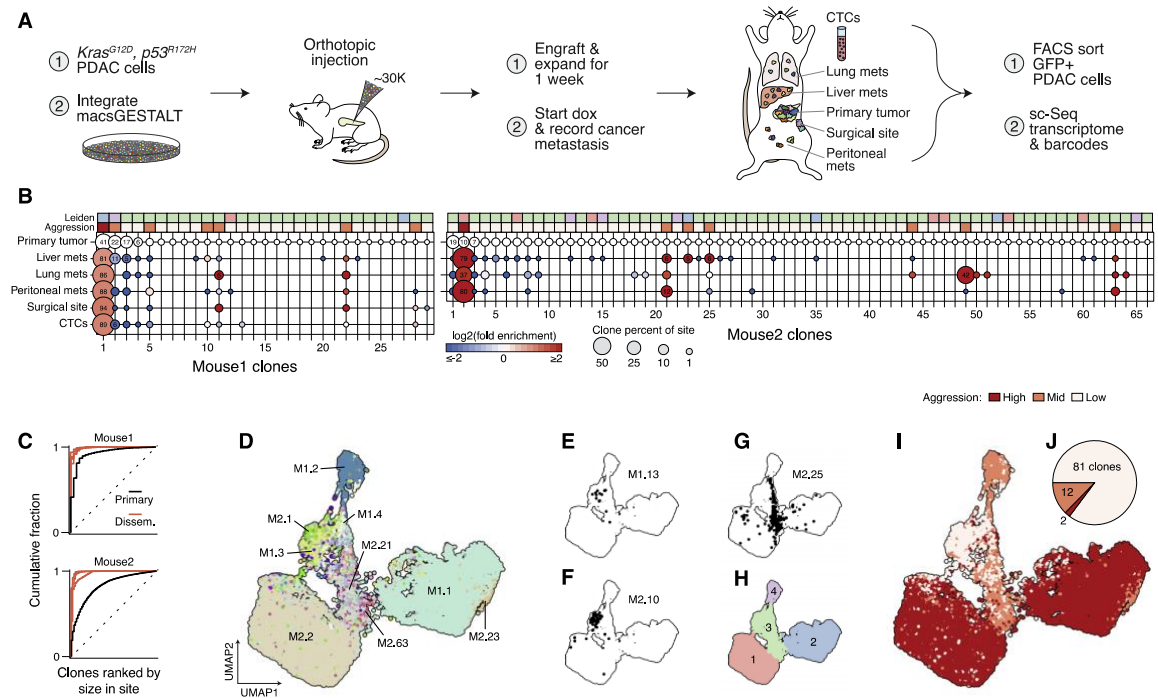
Figure 1



**Figure 1. macsGESTALT for high-resolution lineage tracing**

(A) Genetic components of macsGESTALT. (B) Clone-level information is stored in static barcodes, while subclonal phylogenetic information is dynamically encoded into evolving barcodes via insertions and deletions (indels, blue and red bars) induced by doxycycline. (C) Two example clones from a population with  $n$  clones, each with a random number of integrated barcodes. Evolving barcode edits are encoded and inherited as cells divide. (D) Generation of a macsGESTALT barcoded population of cells and experimental workflow. (E) macsGESTALT analysis workflow. See also Figures S1 and S2.

Figure 2

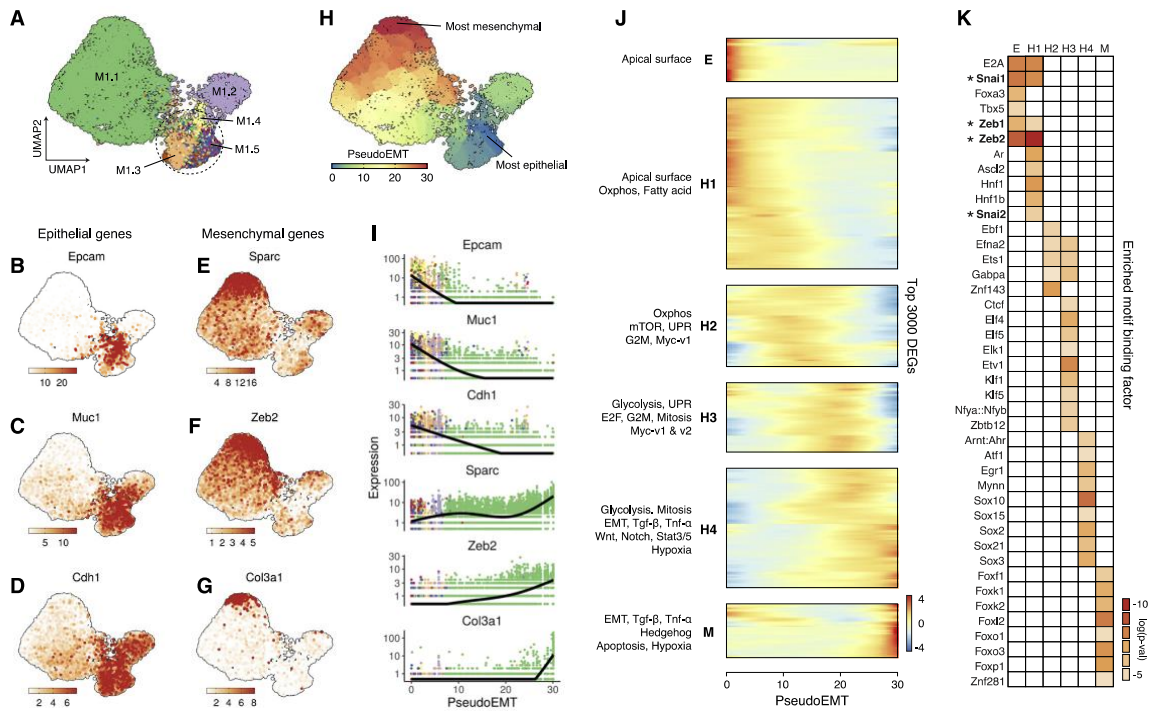


**Figure 2. Most metastases arise from rare, transcriptionally-distinct clones**

(A) Schematic of metastasis lineage tracing model. (B) Clonal reconstruction using static barcodes, where clones are numbered by size in the primary tumor. Percent contribution to each harvest site (circle size) and enrichment compared to the primary tumor (circle color) are visualized. Top annotations show each clone’s Leiden transcriptional cluster and aggression assignments as in (H) and (I), respectively. (C) Cumulative fraction of each CTC clone in each disseminated site (red) and primary tumor (black). Dotted-lines represent the theoretical scenario of perfect clone size equality. (D) UMAP plot of 28,028 single cells containing both lineage and transcriptional information. Cells are colored by clone, with select large clones highlighted (as mouse.clone). (E and F) Two representative non-aggressive clones. (G) A representative clone of medium aggression. (H) Leiden transcriptional clustering of (D). (I) Cells colored by clonal aggression. (J)

Number of non-, mid-, or high-aggression clones of 95 total. See also Figures S3 and S4.

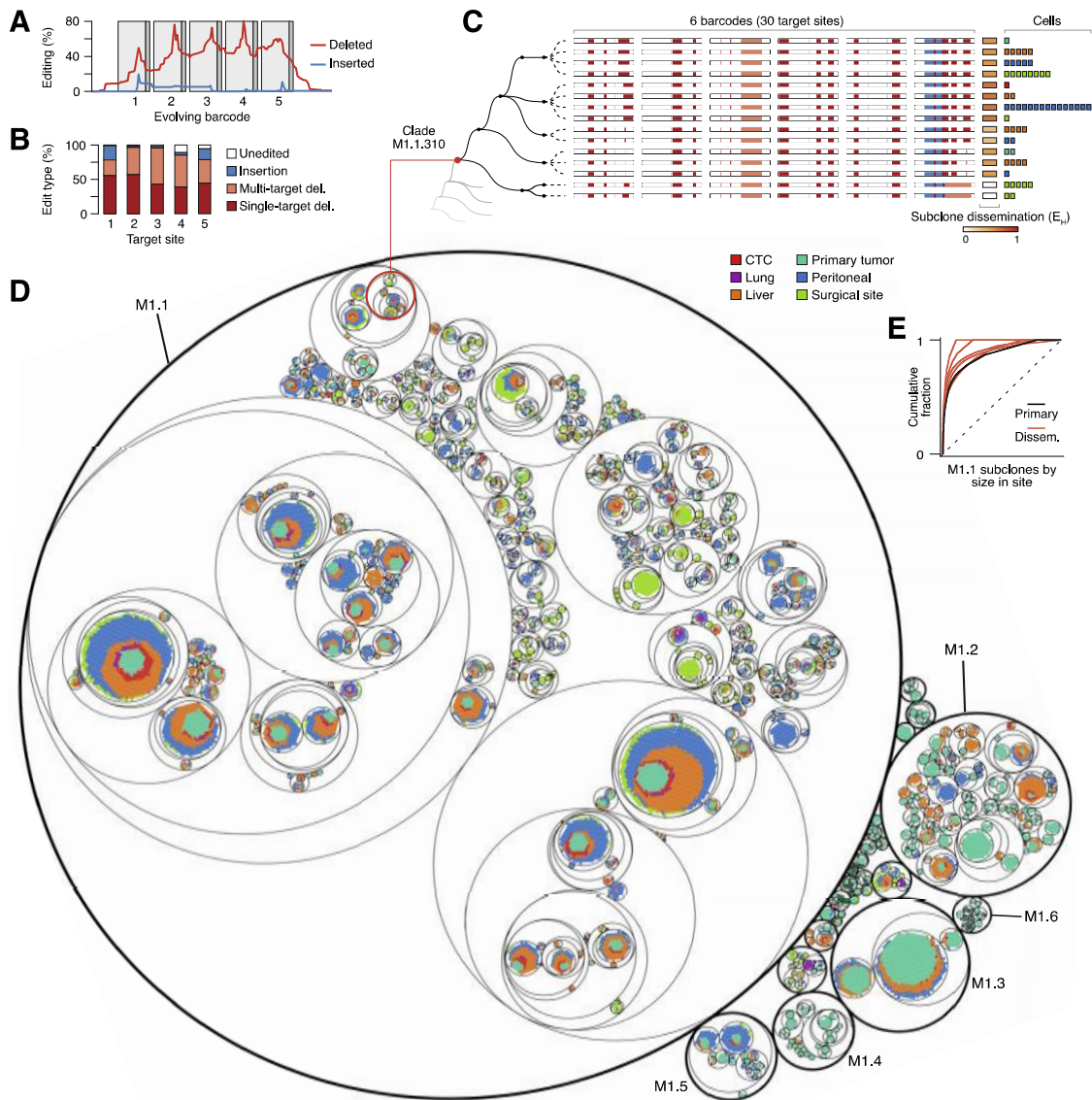
Figure 3



**Figure 3. A transcriptional EMT continuum *in vivo***

(A) UMAP plot of M1, colored by clone, with the five largest clones annotated. Circled region indicates the transcriptional space where smaller, non-aggressive clones reside. (B-G) Expression of canonical epithelial (B-D) and mesenchymal (E-G) markers. (H) Unbiased trajectory inference revealing a pseudotime axis matching EMT (pseudoEMT). (I) Expression of (B-G) plotted along pseudoEMT and colored by clone as in (A). (J) Hierarchical clustering of kinetic curves for the top 3000 differentially expressed genes across pseudoEMT ( $q = 0$ , Moran's  $I > 0.1$ ). Gene clusters are labeled from epithelial [E] to hybrid [H1-H4] to mesenchymal [M] based on expression across pseudoEMT. Geneset analysis using MSigDB Hallmarks for each gene cluster (hypergeometric test,  $p < 0.05$ ). (K) Significantly enriched motifs (hypergeometric test,  $p < 0.05$ ) in promoters for each gene cluster, with canonical EMT master regulators highlighted. See also Figure S5 and Tables S1-S3.

Figure 4

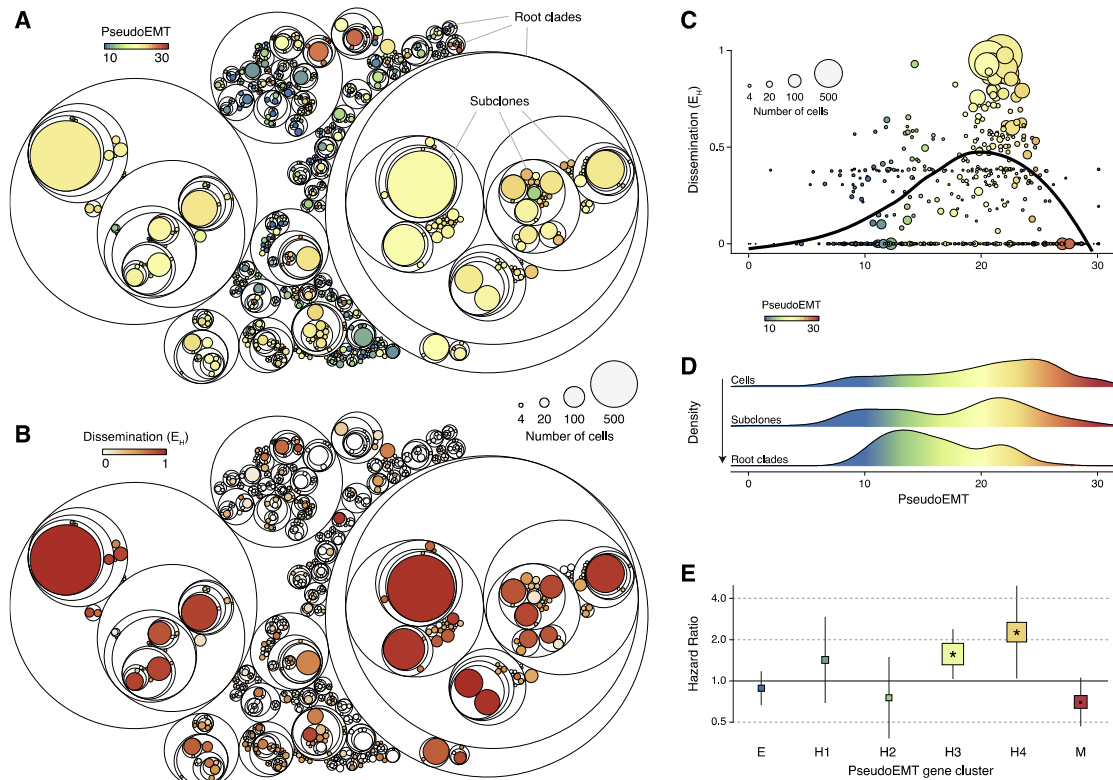


**Figure 4. High-resolution subclonal lineage reconstruction of metastatic cancer**

(A) Percent at which each base is mutated in 76,974 evolving barcodes across both mice. Target site spacers (light grey) and PAMs (dark grey). (B) Edit types observed at each target site. (C) Example phylogenetic reconstruction of a small clade within clone M1.1. Clade M1.1.310 (root node in red) contains 6 distinct subclones composed of 58 cells from 5 different harvest sites. Each cell in this clade has 6 evolving barcodes,

illustrated by white bars with edits colored as in (B). Cells with the same barcode editing pattern are grouped into a subclone (terminal black nodes) and dissemination ( $E_H$ ) is quantified. For each subclone, individual cells are stacked and colored by their harvest site on the far right. (D) Circle packing plot of the full single cell phylogeny of M1, with clade M1.1.310 from (C) circled in red. Outermost circles define clones, with the first 6 clones labeled. Within each clone, nested circles group increasingly related cells. Innermost circles contain cells from reconstructed subclones. Each point represents a single cell, colored by harvest site. (E) Cumulative fraction of each subclone of clone M1.1 in each harvest site. Dotted-line represents perfect subclone-size equality. See also Figure S3.

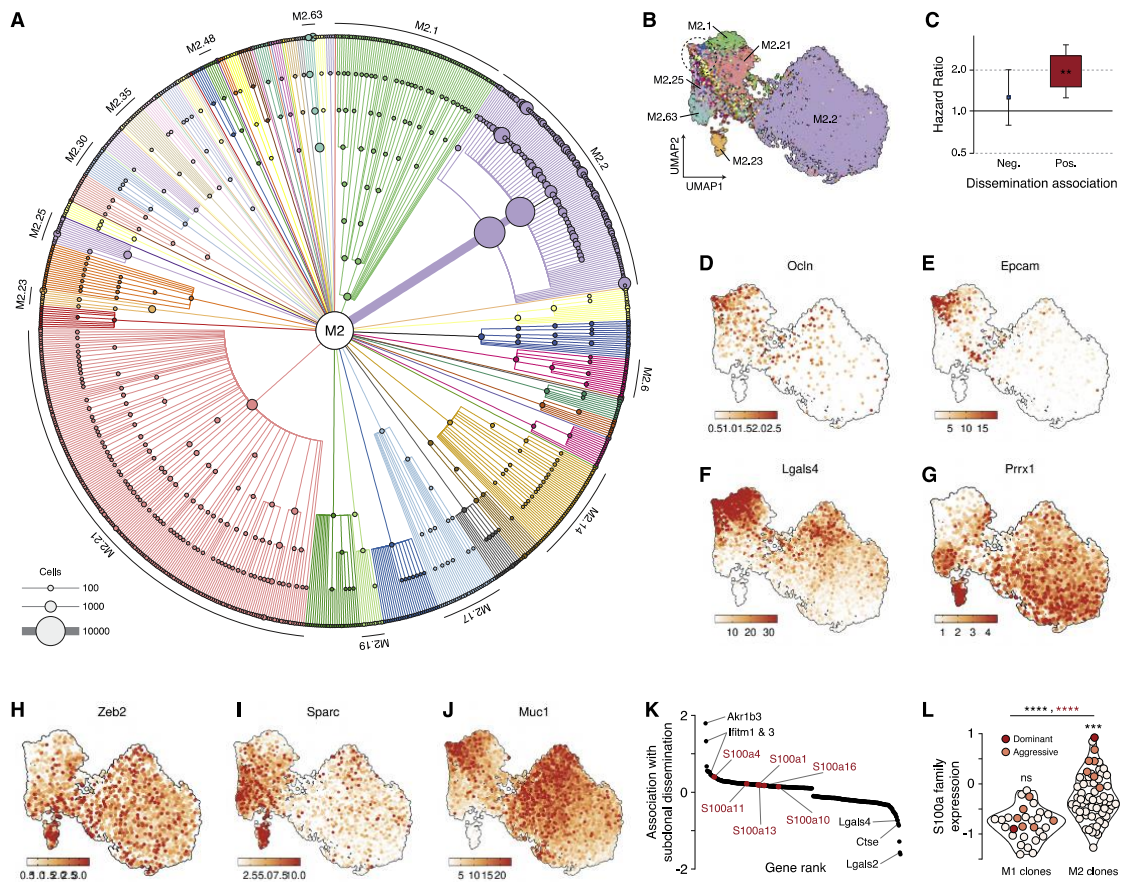
Figure 5



**Figure 5. Peak metastatic aggression corresponds to late-hybrid EMT states**

(A and B) Circle packing plots of the phylogenetic structure of clone M1.1 with subclones colored by mean pseudoEMT (A) and by dissemination score (B). (C) Relationship between metastatic dissemination and pseudoEMT for subclones from (A and B). (D) Density along pseudoEMT of M1.1 cells and their increasingly ancestral (arrow) phylogenetic groupings, examples of which are highlighted in (A). (E) Relationship between PDAC patient survival (TCGA-PAAD,  $n=173$ ) and patient enrichment scores for each pseudoEMT gene cluster using Cox regression analysis, with the hazard ratio for each gene cluster displayed (\*,  $p < 0.05$ , •,  $p < 0.1$ ). Square sizes are inversely proportional to p-value. See also Table S4.

Figure 6



**Figure 6. A complementary process to canonical EMT**

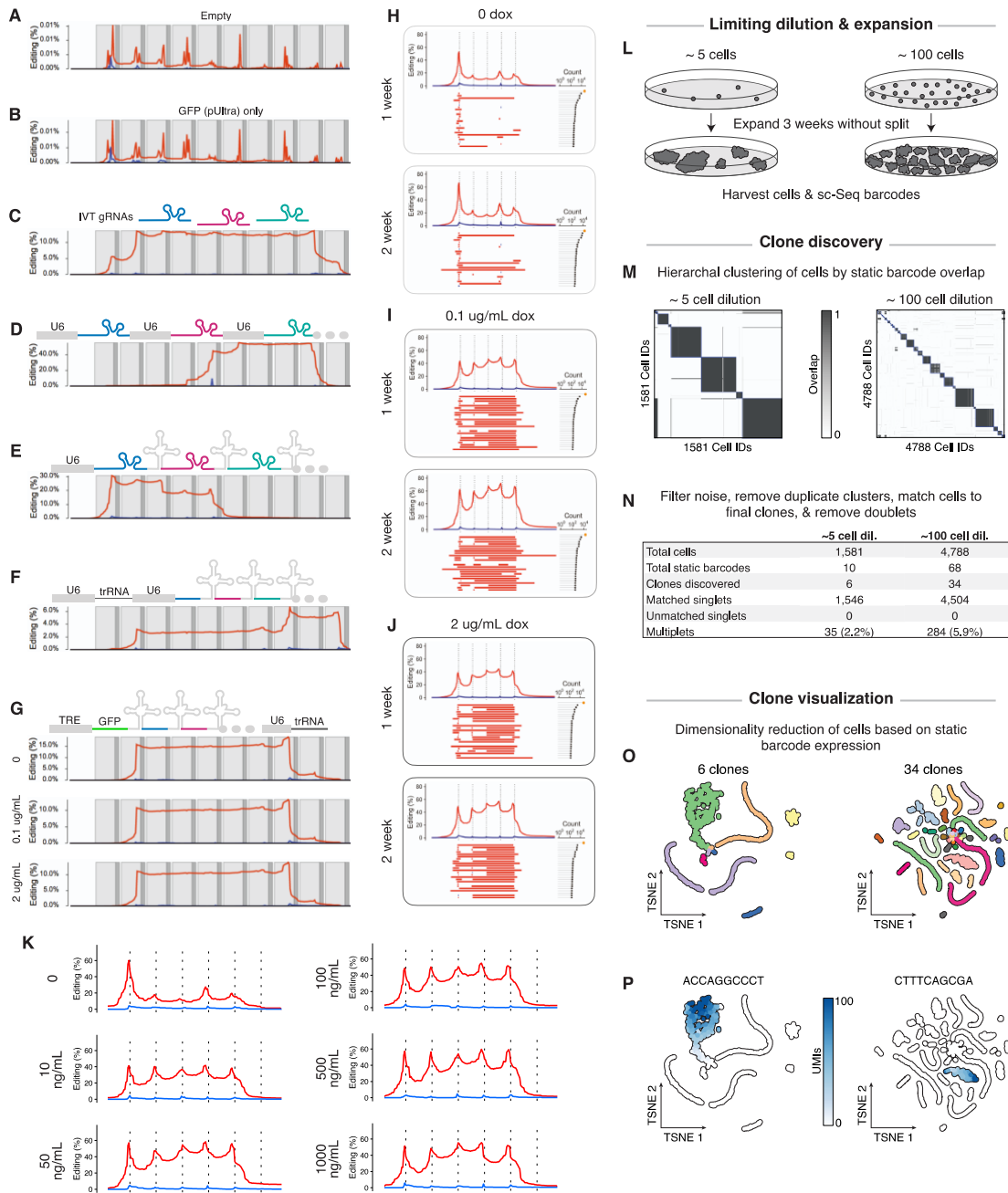
(A) Lineage tree for M2 subclones, where branches and nodes are colored by clone and scaled by the number of cells they relate. (B) UMAP of M2 cells, colored as in (A), with five large, aggressive clones labeled, as well as M2.1 (green), which was the largest clone in the PT but poorly-metastatic. Circled region indicates the transcriptional space where smaller, non-aggressive clones reside. (C) Relationship between PDAC patient survival (TCGA-PAAD, n=173) and enrichment scores for genes associated with subclonal dissemination using Cox regression analysis (\*\*,  $p < 0.01$ ), with the hazard ratio displayed. Square sizes are inversely proportional to p-value. (D-H) Canonical epithelial (D-F) and mesenchymal (G-H) markers. (I and J) Markers with inconsistent



expression patterns in the dominant clone, M2.2. **(K)** Highly expressed genes ranked by association ( $q < 0.05$ ) with subclonal dissemination. **(L)** Aggregated single-cell gene expression of the *S100a* family for each clone, colored by aggression (as defined in Figure 2B) and grouped by mouse. Intramouse comparisons between dominant/aggressive clones versus all others are indicated above each violin. Comparisons between mice for all clones (black) and only dominant/aggressive clones (red) are indicated above the line (Welch's t-test, \*\*\*\*,  $p < 0.0001$ , \*\*\*,  $p < 0.001$ , ns, not significant). See also Figure S6 and Table S5.

# Supplementary figures

## Supplementary figure 1



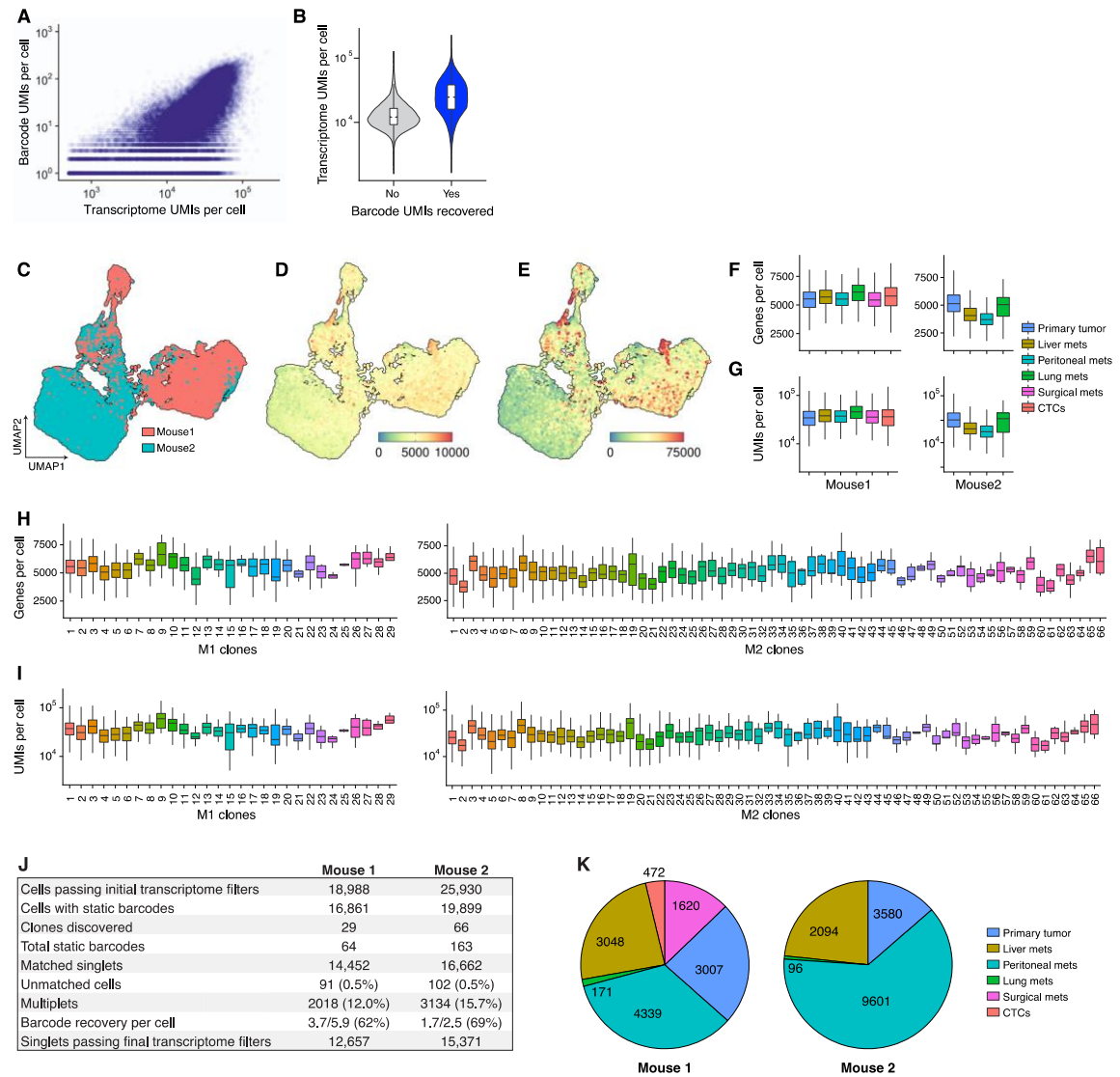
**Figure S1. Designing and validating macsGESTALT, Related to Figure 1**

(A-G) A gRNA array editing screen was performed, where barcoded 293T cells were transfected with constitutive Cas9 vector (px330) and co-transfected with a variety of

controls or gRNA expression formats. Barcode genomic DNA was collected and bulk sequenced one week post-transfection, and for each condition, the percent at which each barcode base was deleted (red) or adjacent to an insertion (blue) is indicated, along with target site spacers (light grey) and PAMs (dark grey). Conditions included: **(A)** No transfection (only Cas9) negative control. **(B)** GFP-only negative control. **(C)** *In vitro* transcribed (IVT) gRNAs positive control. **(D)** Each gRNA placed (targeting sites 5-9) under its own U6 promoter. **(E)** gRNA-tRNA array (targeting sites 1-5) under a U6 promoter (selected for PDAC experiments due to both high editing rate and compact size). **(F)** A split gRNA array with a crRNA-tRNA portion (targeting sites 2-10) and a tracrRNA portion under U6 promoters, where the crRNA portions can complex with the tracrRNA portions when expressed. **(G)** The same array is in (F) but with the crRNA-tRNA array in 3'UTR of a dox-inducible GFP and cultured in three different doses of dox post-transfection for 5 d (this configuration was leaky with no change in editing rate with dox administration). **(H)** Dox-induced macsGESTALT PDAC cells edit evenly across sites and accumulate edits over time. macsGESTALT PDAC cells cultured in dox for one (top) or two (bottom) weeks, and barcodes were bulk DNA sequenced. The percent at which each barcode base was deleted (red) or adjacent to an insertion (blue) is indicated, along with expected cut sites (dotted lines, 3 bp upstream of PAMs). Beneath editing plots, the top 25 most commonly observed alleles are illustrated with the number of observations for each on the right. Leakiness was primarily localized to the first target site, while sites 2-5 remain largely unmutated until dox administration. **(I)** Dox-inducible editing initiates and peaks at low doses in macsGESTALT PDACs. Cells were cultured under six different dosages of dox for 2 weeks and barcodes were bulk DNA sequenced and editing rates plotted. Prior to the start of PDAC editing experiments (H-I), cells experienced 3 weeks of culture time during barcode drug selection, expansion, and

freeze/thawing, during which time background editing from leakiness was possible. **(L-P)** *In vitro* validation of clonal reconstruction and single cell readout of macsGESTALT. **(L)** macsGESTALT PDAC cells were plated at two limiting dilutions of approximately ~5 or ~100 cells, expanded without splitting (even if confluent), and barcodes were scRNA sequenced. **(M)** Potential clones of expanded cells were identified based on static barcode overlap via hierarchical clustering. **(N)** Approximately the expected number of clones were identified for the 5-cell dilution (a 6th cell was unintentionally plated), while the 100-cell dilution retained a smaller fraction of clones likely due to extended culture time under confluence. All cells were successfully matched to a clone and multiplets were explicitly identified (Methods). **(O)** Cells were plotted in tSNE space based on their static barcode expression. Cells clustered in tSNE space consistently with their clonal assignments. **(P)** Two examples illustrating static barcode expression confined to specific clones.

Supplementary figure 2

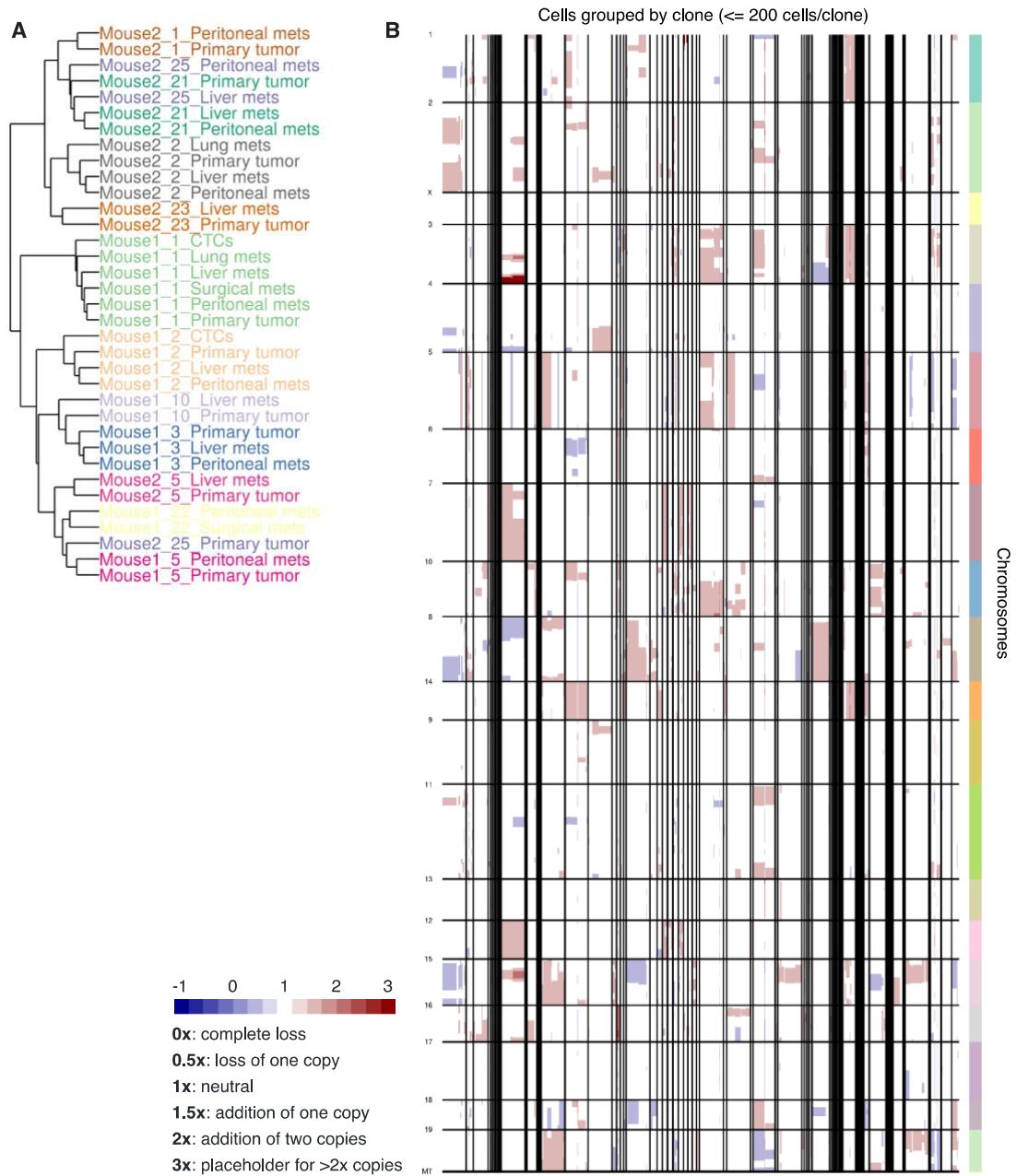


**Figure S2. Summary of single cell transcriptome and barcode recovery in metastasis experiments, Related to Figure 2**

(A) Correlation of barcode UMIs versus transcriptome UMIs (Unique Molecular Identifiers) recovered per cell, for every cell with at least 300 genes captured and at least one barcode UMI captured ( $r = 0.64$ ,  $p < 2.2 \times 10^{-16}$ ). (B) Comparison of transcriptome UMIs recovered per cell for cells grouped by whether or not at least one barcode UMI was recovered (Welch's t-test,  $p < 2.2 \times 10^{-16}$ ). (C) UMAP plot of 28,028 single cells

containing both lineage and transcriptional information passing quality filtering steps (Methods) colored by mouse. **(D)** Colored by genes recovered per cell. **(E)** Colored by transcriptome UMIs recovered per cell. **(F-G)** Genes and UMIs per cell grouped by harvest site for each mouse. **(H-I)** Genes and UMIs per cell grouped by clone for each mouse. **(J)** Summary table of barcode, cell, and clonal recovery for each mouse. **(K)** Number of single cancer cells obtained from each harvest site after filtering.

Supplementary figure 3



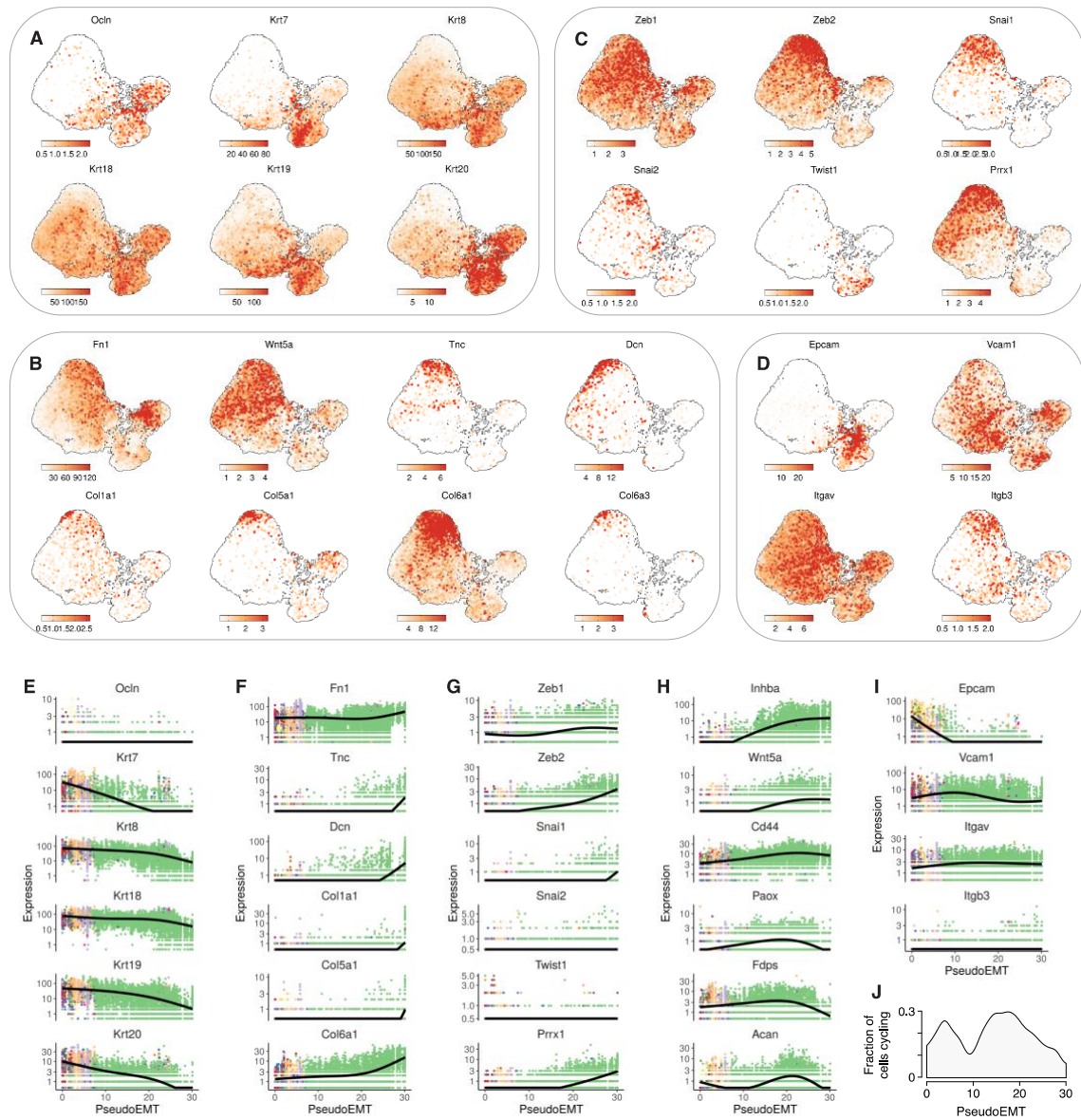
**Figure S3. Clones retain transcriptional identity after metastasizing and CNV among clones, Related to Figure 2**

(A) Cells were analyzed as clone-site pseudobulk samples (i.e. cells from each clone and harvest site combination were aggregated and treated as a bulk sample, see

Methods) and each sample was colored by clone. Only pseudobulk samples with >20 cells were used. Clone-site pseudobulk samples were hierarchically clustered based on whole transcriptome expression. Pseudobulk samples displayed preferential clustering by clone rather than harvest site. **(B)** Genomic copy-number changes among clones. Copy number variation analysis was performed on all 95 clones. Clones with >200 cells were downsampled to 200 cells to perform CNV analysis with InferCNV. Vertical black lines divide clones (many small clones are not visible), and horizontal lines divide chromosomes. Large scale copy number changes are visible between and within clones.



Supplementary figure 4

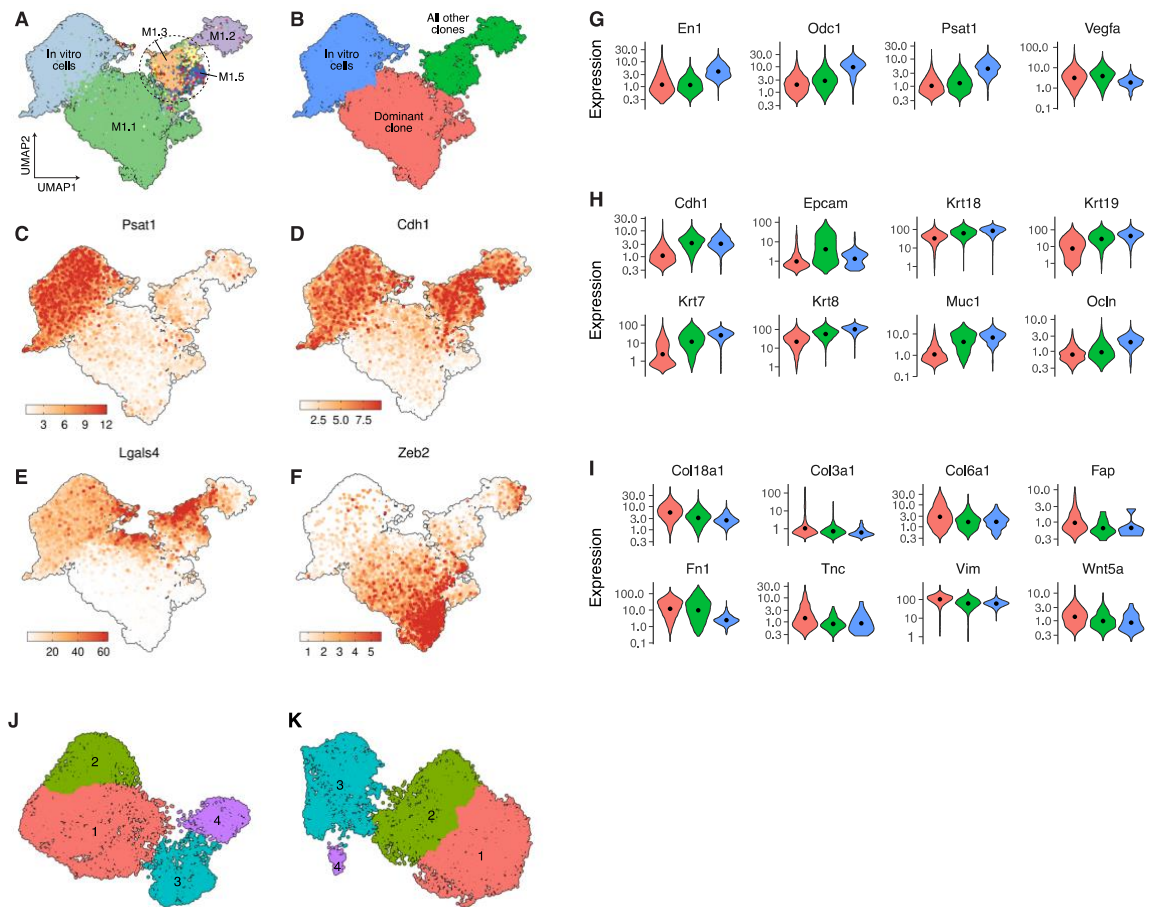


**Figure S4. Gene expression across UMAP space and along pseudoEMT, Related to Figure 3**

(A) Epithelial markers, (B) mesenchymal markers, including extracellular matrix genes, (C) canonical EMT-TFs, and (D) previously used EMT surface markers, expressed in M1 cells. (E) Epithelial markers, (F) extracellular matrix mesenchymal genes, (G) canonical EMT-TFs, (H) selected genes with unusual kinetics, and (I) previously used EMT surface

markers, across pseudoEMT. (**J**) Fraction of cells cycling, i.e. cells in S/G2M cell cycle phase, across pseudoEMT.

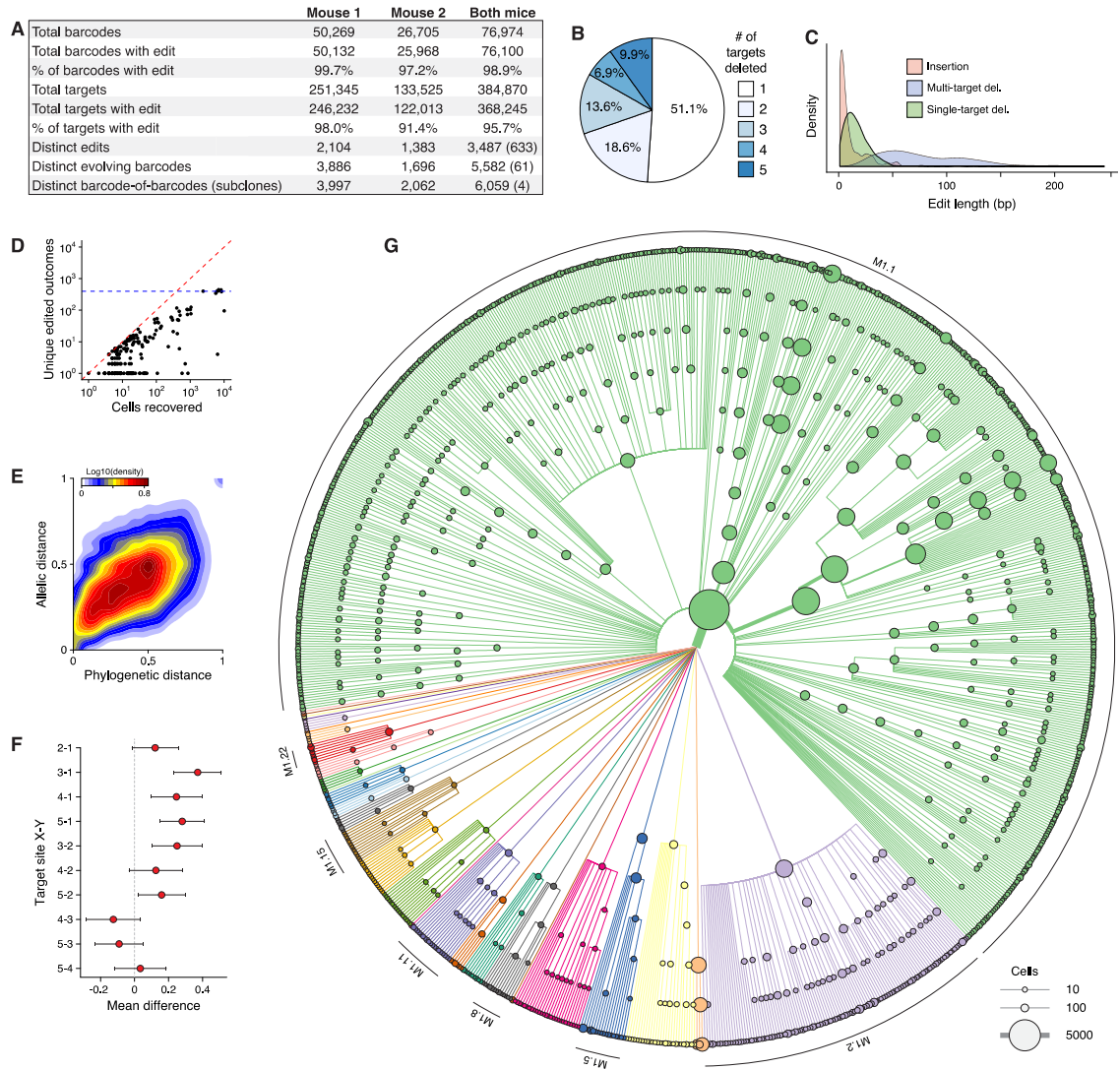
Supplementary figure 5



**Figure S5. Comparison of PDAC cells *in vivo* and *in vitro*, Related to Figure 3**

(A) UMAP of 12,657 *in vivo* M1 single cells and 5,932 *in vitro* cultured single cells, colored by M1 clone or *in vitro* origin. (B) Leiden transcriptional clustering of (A). (C-F) Expression of (C) an *in vitro* cluster top marker gene, (D-E) epithelial markers, and (F) an EMT-TF master regulator. (G-I) Violin plots of (G) *in vitro* cluster top markers, (H) epithelial markers, and (I) mesenchymal markers. (J) Leiden clustering of M1 cells only. (K) Leiden clustering of M2 cells only.

Supplementary figure 6



**Figure S6. Summary of editing and lineage information in metastasis experiments, Related to Figure 4**

(A) Summary table of the number of barcodes/target sites recovered, and the rate at which they were observed to carry a mutation. Additionally, the number of distinct edits, evolving barcodes, and barcode-of-barcodes are displayed. In the last three rows in the last column, the number of overlapping edits, evolving barcodes, and barcode-of-barcodes between the mice is indicated in parentheses. (B) The proportion at which a

deletion impacts 1, 2, 3, 4, or 5 target sites. **(C)** Size distribution for insertions, single-target deletions, and multi-target deletions. **(D)** Visualization of the barcode editing diversity recovered at different cell recovery rates, illustrating a plateauing maximum possible diversity per barcode. For each barcode integrant, the number of cells in which it was recovered versus the number of unique edited outcomes (alleles) detected is plotted. The red dashed line represents the 1:1 maximum diversity scenario, where every cell recovered has a unique edited outcome. The blue dashed line marks 400 unique edited outcomes, i.e. the approximate maximum observed at any cell recovery number. **(E)** Pairwise concordance between phylogenetic distance (distance on reconstructed trees) and barcode allelic distance (the number of edits required to convert between alleles) for all clones in both M1 and M2. **(F)** Mean difference in the first appearance of editing in lineage trees of both mice for all target pairs, with 95% Tukey confidence limits. Editing events were filtered to only include single-target events to avoid confounding from large deletions. **(G)** Lineage tree for M1 subclones, where branches and nodes are colored by clone (as in Figure 3A) and scaled by the number of cells they relate (alternative visualization of the Figure 4D circle packing plot of M1).

## Supplementary tables

All supplementary table files can be found at: <https://doi.org/10.1016/j.ccell.2021.05.005>

### *Supplementary table 1*

*Table S1. Top marker genes for Leiden clusters comparing in vivo and in vitro cells, Related to Figure 3*

### *Supplementary table 2*

Table S2. Differentially expressed genes and MSigDB Hallmark gene sets enriched across pseudoEMT, Related to Figure 3

### *Supplementary table 3*

Table S3. Top marker genes for M1 and M2 Leiden clusters and clones, Related to Figure 2

### *Supplementary table 4*

Table S4. TCGA survival analysis across cancers for each pseudoEMT gene cluster, Related to Figure 5

### *Supplementary table 5*

Table S5. Genes associated with M2 subclonal dissemination and TCGA survival analysis, Related to Figure 6

### *Supplementary table 6*

Table S6. Annotated primer sequences and annealing temperatures used for bulk and single-cell barcode sequencing, Related to STAR Methods

## CHAPTER 3: DISCUSSION

Research into the genetic drivers of metastasis has proven challenging and often unproductive when compared to the highly successful studies characterizing the genetic drivers of tumorigenesis (Kandoth et al. 2013; Hutter and Zenklusen 2018; Esposito, Ganesan, and Kang 2021). Instead epigenetic and transcriptional processes have been hypothesized to potentially play a more consistent role in promoting metastasis. In order to characterize the non-genetic adaptations that may enable metastasis to emerge from the heterogenous morass of cancer, we require tools that can faithfully and precisely identify metastasis-capable subpopulations and simultaneously capture nongenetic cell state information.

Hitherto, tools that can reconstruct tumor population structure have relied on retrospective lineage tracing methods, which are limited by their resolution and often their ability to concurrently attain cell state information. The advent of static barcoding enabled vastly improved clonal labeling diversity than what is possible with retrospective methods confined to natural diversity or prospective fluorescence methods confined by the amount of colors that can be distinguished by microscopy (R. Lu et al. 2011; Kretzschmar and Watt 2012). Static barcoding methods have recently been combined with scRNA-seq readout to allow for capture of precision cell state information (Weinreb et al. 2020). However, static approaches are generally confined to introducing labels *in vitro*, thereby missing any cell state diversity that emerges after this point or *in vivo*, for example after cells are engrafted to form a tumor.

To overcome this technological hurdle to metastasis research, we turned our attention to the field of evolving barcoding, which was just starting to employ CRISPR-Cas9 strategies for barcode mutagenesis, as this thesis work was beginning in late 2016 and early 2017 (McKenna et al. 2016; Frieda et al. 2017; Kalhor, Mali, and Church 2017). Evolving barcoding allowed for the introduction of labels *in vivo* concurrent with the experimental time course of interest, while maintaining the superior labeling diversity enabled by static barcoding methods. Furthermore, evolving barcoding approaches permitted repeated labeling, instead of a single labeling time point as in most static methods.

However until 2021, evolving barcoding methods had generally been confined to studies focusing on *in vitro* validation (Frieda et al. 2017; Kalhor, Mali, and Church 2017; Loveless et al. 2021), normal development (McKenna et al. 2016; Raj et al. 2018; Spanjaard et al. 2018; Alemany et al. 2018; Kalhor et al. 2018; Chan et al. 2019), or adult hematopoiesis (Bowling et al. 2020), usually by injection or transient delivery of lineage tracing components. In order to apply evolving barcoding to study metastasis, we envisioned a compact, easily-integratable system with improved labeling diversity. By merging static barcoding with evolving barcoding, we developed macsGESTALT (Simeonov et al. 2021; Lee and Kang 2021), a flexible, inducible lineage tracer that generates thousands of unique labels both *in vitro* and *in vivo* and that can be readily coupled with scRNA-seq. We applied macsGESTALT to understand pancreatic cancer metastasis and gained critical insight into cancer behavior at the clonal, subclonal, and transcriptomic levels.



Despite using an aggressive genetic model of pancreatic cancer metastasis, we found that most clones do not contribute to metastasis and that only very rare clones contribute significantly, supporting the importance of transcriptional and non-genetic processes ([Hunter et al. 2018](#)). While non-aggressive clones occupied similar transcriptional space, many aggressive clones conversely had distinct transcriptional identities, which they retained even upon dissemination to distant metastatic sites. Among aggressive clones, we found that a single dominant clone drove the overwhelming majority of metastasis across all sites, without apparent organotropism. This lack of apparent organotropism has recently also been observed in models of lung and breast cancer metastasis ([W. Zhang et al. 2021](#); [Quinn et al. 2021](#)). Thereby suggesting that metastatic ability confers a shared ability to metastasize without significant predilection for common sites, at least in models of aggressive cancer. We note our findings were remarkably consistent across mice and suspect that the emergence of rare dominant clones from many non-metastatic clones may be a conserved feature of metastasis in this PDAC model. Additionally, these findings regarding population structure of metastases are strikingly similar to findings in lung cancer metastasis ([Quinn et al. 2021](#)).

Highlighting the limitations of static barcoding approaches in isolation that we discussed previously, extensive clonal bottlenecking obscured lineage information at critical points *in vivo*. By pairing static and evolving barcodes, macsGESTALT overcame these clonal bottlenecking challenges by enabling subclonal reconstruction via the inducible evolving barcodes. In this work, evolving barcodes revealed that growth and dissemination of the dominant clone were driven by rare highly aggressive subclones that were associated with specific EMT transcriptional states. While a wide-range of EMT states existed *in*

*vivo* — from highly epithelial to highly mesenchymal — aggressive subclones exhibited primarily late-hybrid EMT states. These late-hybrid subclones appeared to undergo continuous and aggressive evolutionary selection from a background of predominantly epithelial states. While this process enabled rapid proliferation and metastasis, it also necessitated extensive population bottlenecks, which we note may be a potentially vulnerable or exploitable feature of PDAC metastasis. Further underscoring the therapeutic relevance of our findings, late-hybrid EMT states corresponded with worse overall survival in human PDAC, while epithelial, early-hybrid, or highly mesenchymal states did not, thereby mirroring the rise and fall of metastatic capability across EMT in our model. These findings support studies purporting the importance of EMT in metastasis and drug resistance ([Tsai et al. 2012](#); [Nieto et al. 2016](#); [Fischer et al. 2015](#); [Zheng et al. 2015](#); [Aiello et al. 2017](#)). Recent reports have also described that partial or hybrid EMT states have increased aggressiveness, a concept which we were able to describe with single-cell resolution *in vivo* ([Aiello et al. 2018](#); [Pastushenko et al. 2018](#)). As such, we characterized the EMT spectrum in depth, finding numerous enriched signaling, metabolic, and regulatory features throughout. Amongst these, in late-hybrid EMT states, we observed increased MYC activity and proliferation, as well as potential metabolic rewiring from OXPHOS to glycolysis, which has been implicated in both tumor invasiveness ([Kamarajugadda et al. 2012](#); [J. Lu, Tan, and Cai 2015](#)) and EMT ([Thomson, Balcells, and Cascante 2019](#); [H. Kang et al. 2019](#)).

By exploring the dynamics of a dominant clone driving metastasis in a mouse model of PDAC, we characterized a detailed molecular roadmap of EMT *in vivo* and highlighted one potential path to aggressive metastasis, while noting that *S100* genes may provide a

complementary path as evidenced by their extensive expression propagation, particular amongst aggressive clones. As cancers are notoriously heterogeneous, we anticipate that many different paths to aggressive dissemination likely exist — but promisingly, we find that the late-hybrid EMT states uncovered by macsGESTALT also predict worse survival in a large human patient cohort, suggesting they may be an example of a conserved mechanism. As PDAC has the lowest survival rate of any major cancer (Cancer Facts & Figures 2020), largely due to aggressive, early metastasis present at diagnosis, we hope that our approach will enable future studies to reveal additional processes underlying the highly metastatic nature of PDAC.

Our insights derive from a global, unbiased assessment of metastatic phylogeny and transcription at the single cell level. macsGESTALT enables such investigations by combining static and evolving lineage tracing and achieving high barcode recovery and editing rates, producing rich lineage trees densely annotated with transcriptional information. In this work, we perform lineage tracing of approximately 100 distinct cancer clones across two mice, uncovering both conserved and distinct mechanisms of cancer dissemination. We hope that future work will build on our findings by probing these processes in many mice across multiple cancers. Such studies could eventually exhaustively map the full landscape of cancer heterogeneity to identify the complete repertoire of evolutionary paths leading to metastasis. Functional analyses could then focus on validating these epigenetic and transcriptional avenues. This strategy could be applied to other potentially related aspects of cancer, such as therapy resistance.

While the heterogeneity displayed by cancer across and within individuals, and between tissues of origin, can appear nearly infinite, the evolutionary paths that lead to metastatic competency are likely finite. If we take the analogy of a walled settlement on an unknown alien world with all manner of unknown species that may enter, we would notice that there are three ways to gain entry: above, below, and through the wall. While at first, we do not understand the nature of the various possible alien invaders, by studying them, we can identify those with features that, for example, enable them to fly over the wall. And while the evolutionary adaptations enabling flight on this alien planet may be diverse, they will all likely converge on the displacement of atmospheric gas. We believe that evolving barcoding, alongside rapid advances in single cell multiomics ([Perkel 2021](#)) and signal recording ([Perli, Cui, and Lu 2016](#); [J. Park et al. 2021](#)), will be the tools with which we characterize and stop these alien invaders in the 21st century. And if a species exists that gains access via jumping to great heights rather than by flying, we will be prepared.

## BIBLIOGRAPHY

- Aceto, Nicola, Aditya Bardia, David T. Miyamoto, Maria C. Donaldson, Ben S. Wittner, Joel A. Spencer, Min Yu, et al. 2014. "Circulating Tumor Cell Clusters Are Oligoclonal Precursors of Breast Cancer Metastasis." *Cell* 158 (5): 1110–22.
- Aiello, Nicole M., Thomas Brabletz, Yibin Kang, M. Angela Nieto, Robert A. Weinberg, and Ben Z. Stanger. 2017. "Upholding a Role for EMT in Pancreatic Cancer Metastasis." *Nature* 547 (7661): E7–8.
- Aiello, Nicole M., Ravikanth Maddipati, Robert J. Norgard, David Balli, Jinyang Li, Salina Yuan, Taiji Yamazoe, et al. 2018. "EMT Subtype Influences Epithelial Plasticity and Mode of Cell Migration." *Developmental Cell* 45 (6): 681–95.e4.
- Aiello, Nicole M., Andrew D. Rhim, and Ben Z. Stanger. 2016. "Orthotopic Injection of Pancreatic Cancer Cells." *Cold Spring Harbor Protocols* 2016 (1): db.prot078360.
- Aleman, Anna, Maria Florescu, Chloé S. Baron, Josi Peterson-Maduro, and Alexander van Oudenaarden. 2018. "Whole-Organism Clone Tracing Using Single-Cell Sequencing." *Nature*, March. <https://doi.org/10.1038/nature25969>.
- Altaba, A. Ruiz i., and A. Ruiz i Altaba. 2011. "Hedgehog Signaling and the Gli Code in Stem Cells, Cancer, and Metastases." *Science Signaling*. <https://doi.org/10.1126/scisignal.2002540>.
- Auffinger, B., A. L. Tobias, Y. Han, G. Lee, D. Guo, M. Dey, M. S. Lesniak, and A. U. Ahmed. 2014. "Conversion of Differentiated Cancer Cells into Cancer Stem-like Cells in a Glioblastoma Model after Primary Chemotherapy." *Cell Death and Differentiation* 21 (7): 1119–31.
- Baron, Chloé S., and Alexander van Oudenaarden. 2019. "Unravelling Cellular Relationships during Development and Regeneration Using Genetic Lineage Tracing." *Nature Reviews. Molecular Cell Biology* 20 (12): 753–65.
- Basu, Sayon, Sanith Cheriyaundath, and Avri Ben-Ze'ev. 2018. "Cell-Cell Adhesion: Linking Wnt/ $\beta$ -Catenin Signaling with Partial EMT and Stemness Traits in Tumorigenesis." *F1000Research* 7 (September). <https://doi.org/10.12688/f1000research.15782.1>.
- Beard, Caroline, Konrad Hochedlinger, Kathrin Plath, Anton Wutz, and Rudolf Jaenisch. 2006. "Efficient Method to Generate Single-Copy Transgenic Mice by Site-Specific Integration in Embryonic Stem Cells." *Genesis* 44 (1): 23–28.
- Behjati, Sam, Meritxell Huch, Ruben van Boxtel, Wouter Karthaus, David C. Wedge, Asif U. Tamuri, Inigo Martincorena, et al. 2014. "Genome Sequencing of Normal Cells Reveals Developmental Lineages and Mutational Processes." *Nature* 513 (7518): 422–25.

- Bhang, Hyo-Eun C., David A. Ruddy, Viveksagar Krishnamurthy Radhakrishna, Justina X. Caushi, Rui Zhao, Matthew M. Hims, Angad P. Singh, et al. 2015. "Studying Clonal Dynamics in Response to Cancer Therapy Using High-Complexity Barcoding." *Nature Medicine* 21 (5): 440–48.
- Biezuner, Tamir, Adam Spiro, Ofir Raz, Shiran Amir, Lilach Milo, Rivka Adar, Noa Chapal-Ilani, et al. 2016. "A Generic, Cost-Effective, and Scalable Cell Lineage Analysis Platform." *Genome Research* 26 (11): 1588–99.
- Birkbak, Nicolai J., Aron C. Eklund, Qiyuan Li, Sarah E. McClelland, David Endesfelder, Patrick Tan, Iain B. Tan, Andrea L. Richardson, Zoltan Szallasi, and Charles Swanton. 2011. "Paradoxical Relationship between Chromosomal Instability and Survival Outcome in Cancer." *Cancer Research* 71 (10): 3447–52.
- Boareto, Marcelo, Mohit Kumar Jolly, Aaron Goldman, Mika Pietilä, Sendurai A. Mani, Shiladitya Sengupta, Eshel Ben-Jacob, Herbert Levine, and Jose' N. Onuchic. 2016. "Notch-Jagged Signalling Can Give Rise to Clusters of Cells Exhibiting a Hybrid Epithelial/mesenchymal Phenotype." *Journal of the Royal Society, Interface / the Royal Society* 13 (118). <https://doi.org/10.1098/rsif.2015.1106>.
- Bocci, Federico, Mohit K. Jolly, Satyendra C. Tripathi, Mitzi Aguilar, Samir M. Hanash, Herbert Levine, and José N. Onuchic. 2017. "Numb Prevents a Complete Epithelial-Mesenchymal Transition by Modulating Notch Signalling." *Journal of the Royal Society, Interface / the Royal Society* 14 (136). <https://doi.org/10.1098/rsif.2017.0512>.
- Boumahdi, Soufiane, and Frederic J. de Sauvage. 2020. "The Great Escape: Tumour Cell Plasticity in Resistance to Targeted Therapy." *Nature Reviews. Drug Discovery* 19 (1): 39–56.
- Bowling, Sarah, Duluxan Sritharan, Fernando G. Osorio, Maximilian Nguyen, Priscilla Cheung, Alejo Rodriguez-Fraticelli, Sachin Patel, et al. 2020. "An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells." *Cell* 181 (6): 1410–22.e27.
- Brastianos, Priscilla K., Scott L. Carter, Sandro Santagata, Daniel P. Cahill, Amaro Taylor-Weiner, Robert T. Jones, Eliezer M. Van Allen, et al. 2015. "Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets." *Cancer Discovery* 5 (11): 1164–77.
- Bresnick, Anne R., David J. Weber, and Danna B. Zimmer. 2015. "S100 Proteins in Cancer." *Nature Reviews. Cancer* 15 (2): 96–109.
- Calon, Alexandre, Elisa Espinet, Sergio Palomo-Ponce, Daniele V. F. Tauriello, Mar Iglesias, María Virtudes Céspedes, Marta Sevillano, et al. 2012. "Dependency of Colorectal Cancer on a TGF- $\beta$ -Driven Program in Stromal Cells for Metastasis Initiation." *Cancer Cell* 22 (5): 571–84.

- Campbell, Peter J., Shinichi Yachida, Laura J. Mudie, Philip J. Stephens, Erin D. Pleasance, Lucy A. Stebbings, Laura A. Morsberger, et al. 2010. "The Patterns and Dynamics of Genomic Instability in Metastatic Pancreatic Cancer." *Nature* 467 (7319): 1109–13.
- Cancer Action Network. 2020. "The Costs of Cancer, 2020 Edition." American Cancer Society. 2020.  
<https://www.fightcancer.org/sites/default/files/National%20Documents/Costs-of-Cancer-2020-10222020.pdf>.
- Cancer Facts & Figures. 2020. "Cancer Facts & Figures 2020."  
<https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf>.
- Cancer Genome Atlas Network. 2012. "Comprehensive Molecular Characterization of Human Colon and Rectal Cancer." *Nature* 487 (7407): 330–37.
- Cancer Genome Atlas Research Network. 2017. "Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma." *Cancer Cell* 32 (2): 185–203.e13.
- Cao, Jian, Lizhen Wu, Shang-Min Zhang, Min Lu, William K. C. Cheung, Wesley Cai, Molly Gale, Qi Xu, and Qin Yan. 2016. "An Easy and Efficient Inducible CRISPR/Cas9 Platform with Improved Specificity for Multiple Gene Targeting." *Nucleic Acids Research* 44 (19): e149.
- Cao, Junyue, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, et al. 2019. "The Single-Cell Transcriptional Landscape of Mammalian Organogenesis." *Nature* 566 (7745): 496–502.
- Carter, Scott L., Aron C. Eklund, Isaac S. Kohane, Lyndsay N. Harris, and Zoltan Szallasi. 2006. "A Signature of Chromosomal Instability Inferred from Gene Expression Profiles Predicts Clinical Outcome in Multiple Human Cancers." *Nature Genetics* 38 (9): 1043–48.
- Casasent, Anna K., Aislyn Schalck, Ruli Gao, Emi Sei, Annalyssa Long, William Pangburn, Tod Casasent, Funda Meric-Bernstam, Mary E. Edgerton, and Nicholas E. Navin. 2018. "Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing." *Cell* 172 (1-2): 205–17.e12.
- Chalfie, M., Y. Tu, G. Euskirchen, W. W. Ward, and D. C. Prasher. 1994. "Green Fluorescent Protein as a Marker for Gene Expression." *Science* 263 (5148): 802–5.
- Chan, Michelle M., Zachary D. Smith, Stefanie Grosswendt, Helene Kretzmer, Thomas M. Norman, Britt Adamson, Marco Jost, et al. 2019. "Molecular Recording of Mammalian Embryogenesis." *Nature* 570 (7759): 77–82.
- Chow, Ke-Huan K., Mark W. Budde, Alejandro A. Granados, Maria Cabrera, Shinae Yoon, Soomin Cho, Ting-Hao Huang, et al. 2021. "Imaging Cell Lineage with a

- Synthetic Digital Recording System." *Science* 372 (6538).  
<https://doi.org/10.1126/science.abb3099>.
- Colak, Selcuk, and Peter ten Dijke. 2017. "Targeting TGF- $\beta$  Signaling in Cancer." *Trends in Cancer*. <https://doi.org/10.1016/j.trecan.2016.11.008>.
- Coorens, Tim H. H., Luiza Moore, Philip S. Robinson, Rashesh Sanghvi, Joseph Christopher, James Hewinson, Moritz J. Przybilla, et al. 2021. "Extensive Phylogenies of Human Development Inferred from Somatic Mutations." *Nature* 597 (7876): 387–92.
- Davis, B. W., R. D. Gelber, A. Goldhirsch, W. H. Hartmann, G. W. Locher, R. Reed, R. Golouh, J. Säve-Söderbergh, L. Holloway, and I. Russell. 1986. "Prognostic Significance of Tumor Grade in Clinical Trials of Adjuvant Therapy for Breast Cancer with Axillary Lymph Node Metastasis." *Cancer* 58 (12): 2662–70.
- Davoli, Teresa, Hajime Uno, Eric C. Wooten, and Stephen J. Elledge. 2017. "Tumor Aneuploidy Correlates with Markers of Immune Evasion and with Reduced Response to Immunotherapy." *Science* 355 (6322).  
<https://doi.org/10.1126/science.aaf8399>.
- Dijk, David van, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J. Carr, Cassandra Burdziak, et al. 2018. "Recovering Gene Interactions from Single-Cell Data Using Data Diffusion." *Cell* 174 (3): 716–29.e27.
- DiMeo, Theresa A., Kristen Anderson, Pushkar Phadke, Cheng Fan, Charles M. Perou, Steven Naber, and Charlotte Kuperwasser. 2009. "A Novel Lung Metastasis Signature Links Wnt Signaling with Cancer Cell Self-Renewal and Epithelial-Mesenchymal Transition in Basal-like Breast Cancer." *Cancer Research* 69 (13): 5364–73.
- Domingo-Domenech, Josep, Samuel J. Vidal, Veronica Rodriguez-Bravo, Mireia Castillo-Martin, S. Aidan Quinn, Ruth Rodriguez-Barrueco, Dennis M. Bonal, et al. 2012. "Suppression of Acquired Docetaxel Resistance in Prostate Cancer through Depletion of Notch- and Hedgehog-Dependent Tumor-Initiating Cells." *Cancer Cell*.  
<https://doi.org/10.1016/j.ccr.2012.07.016>.
- Echeverria, Gloria V., Emily Powell, Sahil Seth, Zhongqi Ge, Alessandro Carugo, Christopher Bristow, Michael Peoples, et al. 2018. "High-Resolution Clonal Mapping of Multi-Organ Metastasis in Triple Negative Breast Cancer." *Nature Communications* 9 (1): 5079.
- El-Kebir, Mohammed, Layla Oesper, Hannah Acheson-Field, and Benjamin J. Raphael. 2015. "Reconstruction of Clonal Trees and Tumor Composition from Multi-Sample Sequencing Data." *Bioinformatics* 31 (12): i62–70.
- Esposito, Mark, Shridar Ganesan, and Yibin Kang. 2021. "Emerging Strategies for Treating Metastasis." *Nature Cancer* 2 (3): 258–70.



- Esposito, Mark, Nandini Mondal, Todd M. Greco, Yong Wei, Chiara Spadazzi, Song-Chang Lin, Hanqiu Zheng, et al. 2019. "Bone Vascular Niche E-Selectin Induces Mesenchymal–epithelial Transition and Wnt Activation in Cancer Cells to Promote Bone Metastasis." *Nature Cell Biology* 21 (5): 627–39.
- Fei, Fei, Jie Qu, Mingqing Zhang, Yuwei Li, and Shiwu Zhang. 2017. "S100A4 in Cancer Progression and Metastasis: A Systematic Review." *Oncotarget* 8 (42): 73219–39.
- Felsenstein, Joseph. 1989. "PHYLIP - Phylogeny Inference Package (Version 3.2)." *Cladistics: The International Journal of the Willi Hennig Society* 5: 164–66.
- Fischer, Kari R., Anna Durrans, Sharrell Lee, Jianting Sheng, Fuhai Li, Stephen T. C. Wong, Hyejin Choi, et al. 2015. "Epithelial-to-Mesenchymal Transition Is Not Required for Lung Metastasis but Contributes to Chemoresistance." *Nature* 527 (7579): 472–76.
- Foroutan, Momeneh, Dharmesh D. Bhuvra, Ruqian Lyu, Kristy Horan, Joseph Cursons, and Melissa J. Davis. 2018. "Single Sample Scoring of Molecular Phenotypes." *BMC Bioinformatics* 19 (1): 404.
- Frieda, Kirsten L., James M. Linton, Sahand Hormoz, Joonhyuk Choi, Ke-Huan K. Chow, Zakary S. Singer, Mark W. Budde, Michael B. Elowitz, and Long Cai. 2017. "Synthetic Recording and in Situ Readout of Lineage Information in Single Cells." *Nature* 541 (7635): 107–11.
- Gabay, Meital, Yulin Li, and Dean W. Felsher. 2014. "MYC Activation Is a Hallmark of Cancer Initiation and Maintenance." *Cold Spring Harbor Perspectives in Medicine* 4 (6). <https://doi.org/10.1101/cshperspect.a014241>.
- Grimm, Daniela, Johann Bauer, Petra Wise, Marcus Krüger, Ulf Simonsen, Markus Wehland, Manfred Infanger, and Thomas J. Corydon. 2019. "The Role of SOX Family Members in Solid Tumours and Metastasis." *Seminars in Cancer Biology*, March. <https://doi.org/10.1016/j.semcancer.2019.03.004>.
- Ha, Gavin, Andrew Roth, Jaswinder Khattri, Julie Ho, Damian Yap, Leah M. Prentice, Nataliya Melnyk, et al. 2014. "TITAN: Inference of Copy Number Architectures in Clonal Cell Populations from Tumor Whole-Genome Sequence Data." *Genome Research* 24 (11): 1881–93.
- Hajirasouliha, Iman, Ahmad Mahmood, and Benjamin J. Raphael. 2014. "A Combinatorial Approach for Analyzing Intra-Tumor Heterogeneity from High-Throughput Sequencing Data." *Bioinformatics* 30 (12): i78–86.
- Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. "Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities." *Molecular Cell* 38 (4): 576–89.

- Hingorani, Sunil R., Lifu Wang, Asha S. Multani, Chelsea Combs, Therese B. Deramaudt, Ralph H. Hruban, Anil K. Rustgi, Sandy Chang, and David A. Tuveson. 2005. "Trp53R172H and KrasG12D Cooperate to Promote Chromosomal Instability and Widely Metastatic Pancreatic Ductal Adenocarcinoma in Mice." *Cancer Cell* 7 (5): 469–83.
- Hong, Matthew K. H., Geoff Macintyre, David C. Wedge, Peter Van Loo, Keval Patel, Sebastian Lunke, Ludmil B. Alexandrov, et al. 2015. "Tracking the Origins and Drivers of Subclonal Metastatic Expansion in Prostate Cancer." *Nature Communications* 6 (April): 6605.
- Hong, Tian, Kazuhide Watanabe, Catherine Ha Ta, Alvaro Villarreal-Ponce, Qing Nie, and Xing Dai. 2015. "An Ovol2-Zeb1 Mutual Inhibitory Circuit Governs Bidirectional and Multi-Step Transition between Epithelial and Mesenchymal States." *PLoS Computational Biology* 11 (11): e1004569.
- Hsu, Tien, Maria Trojanowska, and Dennis K. Watson. 2004. "Ets Proteins in Biological Control and Cancer." *Journal of Cellular Biochemistry* 91 (5): 896–903.
- Hunter, Kent W., Ruhul Amin, Sarah Deasy, Ngoc-Han Ha, and Lalage Wakefield. 2018. "Genetic Insights into the Morass of Metastatic Heterogeneity." *Nature Reviews. Cancer* 18 (4): 211–23.
- Hutter, Carolyn, and Jean Claude Zenklusen. 2018. "The Cancer Genome Atlas: Creating Lasting Value beyond Its Data." *Cell* 173 (2): 283–85.
- Jamal-Hanjani, M., R. A'Hern, N. J. Birkbak, P. Gorman, E. Grönroos, S. Ngang, P. Nicola, et al. 2015. "Extreme Chromosomal Instability Forecasts Improved Outcome in ER-Negative Breast Cancer: A Prospective Validation Cohort Study from the TACT Trial." *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 26 (7): 1340–46.
- Jolly, Mohit Kumar, Marcelo Boareto, Bin Huang, Dongya Jia, Mingyang Lu, Eshel Ben-Jacob, José N. Onuchic, and Herbert Levine. 2015. "Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis." *Frontiers in Oncology* 5 (July): 155.
- Ju, Young Seok, Inigo Martincorena, Moritz Gerstung, Mia Petljak, Ludmil B. Alexandrov, Raheleh Rahbari, David C. Wedge, et al. 2017. "Somatic Mutations Reveal Asymmetric Cellular Dynamics in the Early Human Embryo." *Nature* 543 (7647): 714–18.
- Kalhor, Reza, Kian Kalhor, Leo Mejia, Kathleen Leeper, Amanda Graveline, Prashant Mali, and George M. Church. 2018. "Developmental Barcoding of Whole Mouse via Homing CRISPR." *Science*, August. <https://doi.org/10.1126/science.aat9804>.
- Kalhor, Reza, Prashant Mali, and George M. Church. 2017. "Rapidly Evolving Homing CRISPR Barcodes." *Nature Methods* 14 (2): 195–200.

- Kamarajugadda, Sushama, Lauren Stemboroski, Qingsong Cai, Nicholas E. Simpson, Sushrusha Nayak, Ming Tan, and Jianrong Lu. 2012. "Glucose Oxidation Modulates Anoikis and Tumor Metastasis." *Molecular and Cellular Biology* 32 (10): 1893–1907.
- Kandoth, Cyriac, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, et al. 2013. "Mutational Landscape and Significance across 12 Major Cancer Types." *Nature* 502 (7471): 333–39.
- Kang, Eunju, Xinjian Wang, Rebecca Tippner-Hedges, Hong Ma, Clifford D. L. Folmes, Nuria Marti Gutierrez, Yeonmi Lee, et al. 2016. "Age-Related Accumulation of Somatic Mitochondrial DNA Mutations in Adult-Derived Human iPSCs." *Cell Stem Cell* 18 (5): 625–36.
- Kang, Hyunkoo, Hyunwoo Kim, Sungmin Lee, Hyesook Youn, and Buhyun Youn. 2019. "Role of Metabolic Reprogramming in Epithelial-Mesenchymal Transition (EMT)." *International Journal of Molecular Sciences* 20 (8).  
<https://doi.org/10.3390/ijms20082042>.
- Kawakubo, Tomoyo, Kuniaki Okamoto, Jun-Ichi Iwata, Masashi Shin, Yoshiko Okamoto, Atsushi Yasukochi, Keiichi I. Nakayama, Tomoko Kadowaki, Takayuki Tsukuba, and Kenji Yamamoto. 2007. "Cathepsin E Prevents Tumor Growth and Metastasis by Catalyzing the Proteolytic Release of Soluble TRAIL from Tumor Cell Surface." *Cancer Research* 67 (22): 10869–78.
- Kebschull, Justus M., and Anthony M. Zador. 2018. "Cellular Barcoding: Lineage Tracing, Screening and beyond." *Nature Methods* 15 (11): 871–79.
- Kester, Lennart, and Alexander van Oudenaarden. 2018. "Single-Cell Transcriptomics Meets Lineage Tracing." *Cell Stem Cell*, May.  
<https://doi.org/10.1016/j.stem.2018.04.014>.
- Kim, Kwonseop, Zifan Lu, and Elizabeth D. Hay. 2002. "Direct Evidence for a Role of Beta-catenin/LEF-1 Signaling Pathway in Induction of EMT." *Cell Biology International* 26 (5): 463–76.
- Kretzschmar, Kai, and Fiona M. Watt. 2012. "Lineage Tracing." *Cell* 148 (1-2): 33–45.
- Lambert, Arthur W., Diwakar R. Pattabiraman, and Robert A. Weinberg. 2017. "Emerging Biological Principles of Metastasis." *Cell* 168 (4): 670–91.
- Lambert, Arthur W., and Robert A. Weinberg. 2021. "Linking EMT Programmes to Normal and Neoplastic Epithelial Stem Cells." *Nature Reviews. Cancer* 21 (5): 325–38.
- Lan, Xiaoyang, David J. Jörg, Florence M. G. Cavalli, Laura M. Richards, Long V. Nguyen, Robert J. Vanner, Paul Guilhamon, et al. 2017. "Fate Mapping of Human Glioblastoma Reveals an Invariant Stem Cell Hierarchy." *Nature*, August.  
<https://doi.org/10.1038/nature23666>.

- Lareau, Caleb A., Leif S. Ludwig, Christoph Muus, Satyen H. Gohil, Tongtong Zhao, Zachary Chiang, Karin Pelka, et al. 2021. "Massively Parallel Single-Cell Mitochondrial DNA Genotyping and Chromatin Profiling." *Nature Biotechnology* 39 (4): 451–61.
- Lee, Eunmi, and Yibin Kang. 2021. "Lineage Tracing Reveals Metastatic Dynamics." *Cancer Cell*, June. <https://doi.org/10.1016/j.ccell.2021.06.005>.
- Liau, Brian B., Cem Sievers, Laura K. Donohue, Shawn M. Gillespie, William A. Flavahan, Tyler E. Miller, Andrew S. Venteicher, et al. 2017. "Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance." *Cell Stem Cell* 20 (2): 233–46.e7.
- Li, Jinyang, Katelyn T. Byrne, Fangxue Yan, Taiji Yamazoe, Zeyu Chen, Timour Baslan, Lee P. Richman, et al. 2018. "Tumor Cell-Intrinsic Factors Underlie Heterogeneity of Immune Cell Infiltration and Response to Immunotherapy." *Immunity* 49 (1): 178–93.e7.
- Li, Ruoyan, Lin Di, Jie Li, Wenyi Fan, Yachen Liu, Wenjia Guo, Weiling Liu, et al. 2021. "A Body Map of Somatic Mutagenesis in Morphologically Normal Human Tissues." *Nature* 597 (7876): 398–403.
- Liu, Reng-Yun, Yuanyuan Zeng, Zhe Lei, Longqiang Wang, Haiping Yang, Zeyi Liu, Jun Zhao, and Hong-Tao Zhang. 2014. "JAK/STAT3 Signaling Is Required for TGF- $\beta$ -Induced Epithelial-Mesenchymal Transition in Lung Cancer Cells." *International Journal of Oncology* 44 (5): 1643–51.
- Liu, Xi, Lu Chen, Yinghui Fan, Yi Hong, Xiaoqun Yang, Yao Li, Jianlei Lu, et al. 2019. "IFITM3 Promotes Bone Metastasis of Prostate Cancer Cells by Mediating Activation of the TGF- $\beta$  Signaling Pathway." *Cell Death & Disease* 10 (7): 517.
- Livet, Jean, Tammy A. Weissman, Hyuno Kang, Ryan W. Draft, Ju Lu, Robyn A. Bennis, Joshua R. Sanes, and Jeff W. Lichtman. 2007. "Transgenic Strategies for Combinatorial Expression of Fluorescent Proteins in the Nervous System." *Nature* 450 (7166): 56–62.
- Lou, Emil, Sho Fujisawa, Alexei Morozov, Afsar Barlas, Yevgeniy Romin, Yildirim Dogan, Sepideh Gholami, André L. Moreira, Katia Manova-Todorova, and Malcolm A. S. Moore. 2012. "Tunneling Nanotubes Provide a Unique Conduit for Intercellular Transfer of Cellular Contents in Human Malignant Pleural Mesothelioma." *PLoS One* 7 (3): e33093.
- Loveless, Theresa B., Joseph H. Grotts, Mason W. Schechter, Elmira Forouzmand, Courtney K. Carlson, Bijan S. Agahi, Guohao Liang, et al. 2021. "Lineage Tracing and Analog Recording in Mammalian Cells by Single-Site DNA Writing." *Nature Chemical Biology* 17 (6): 739–47.
- Ludwig, Leif S., Caleb A. Lareau, Jacob C. Ulirsch, Elena Christian, Christoph Muus, Lauren H. Li, Karin Pelka, et al. 2019. "Lineage Tracing in Humans Enabled by

- Mitochondrial Mutations and Single-Cell Genomics.” *Cell*.  
<https://doi.org/10.1016/j.cell.2019.01.022>.
- Lu, Jianrong, Ming Tan, and Qingsong Cai. 2015. “The Warburg Effect in Tumor Progression: Mitochondrial Oxidative Metabolism as an Anti-Metastasis Mechanism.” *Cancer Letters* 356 (2 Pt A): 156–64.
- Lu, Mingyang, Mohit Kumar Jolly, Herbert Levine, José N. Onuchic, and Eshel Ben-Jacob. 2013. “MicroRNA-Based Regulation of Epithelial-Hybrid-Mesenchymal Fate Determination.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (45): 18144–49.
- Lu, Rong, Norma F. Neff, Stephen R. Quake, and Irving L. Weissman. 2011. “Tracking Single Hematopoietic Stem Cells in Vivo Using High-Throughput Sequencing in Conjunction with Viral Genetic Barcoding.” *Nature Biotechnology* 29 (10): 928–33.
- Maddipati, Ravikanth, and Ben Z. Stanger. 2015. “Pancreatic Cancer Metastases Harbor Evidence of Polyclonality.” *Cancer Discovery* 5 (10): 1086–97.
- Makohon-Moore, Alvin P., Ming Zhang, Johannes G. Reiter, Ivana Bozic, Benjamin Allen, Deepanjan Kundu, Krishnendu Chatterjee, et al. 2017. “Limited Heterogeneity of Known Driver Gene Mutations among the Metastases of Individual Patients with Pancreatic Cancer.” *Nature Genetics* 49 (3): 358–66.
- Margonis, Georgios Antonios, Yuhree Kim, Gaya Spolverato, Aslam Ejaz, Rohan Gupta, David Cosgrove, Robert Anders, Georgios Karagkounis, Michael A. Choti, and Timothy M. Pawlik. 2015. “Association Between Specific Mutations in KRAS Codon 12 and Colorectal Liver Metastasis.” *JAMA Surgery* 150 (8): 722–29.
- Marine, Jean-Christophe, Sarah-Jane Dawson, and Mark A. Dawson. 2020. “Non-Genetic Mechanisms of Therapeutic Resistance in Cancer.” *Nature Reviews. Cancer* 20 (12): 743–56.
- Marshall, Eliot. 2011. “Cancer Research and the \$90 Billion Metaphor.” *Science* 331 (6024): 1540–41.
- McCreery, Melissa Q., Kyle D. Halliwill, Douglas Chin, Reyno Delrosario, Gillian Hirst, Peter Vuong, Kuang-Yu Jen, James Hewinson, David J. Adams, and Allan Balmain. 2015. “Evolution of Metastasis Revealed by Mutational Landscapes of Chemically Induced Skin Cancers.” *Nature Medicine* 21 (12): 1514–20.
- McFaline-Figueroa, José L., Andrew J. Hill, Xiaojie Qiu, Dana Jackson, Jay Shendure, and Cole Trapnell. 2019. “A Pooled Single-Cell Genetic Screen Identifies Regulatory Checkpoints in the Continuum of the Epithelial-to-Mesenchymal Transition.” *Nature Genetics* 51 (9): 1389–98.
- McGranahan, Nicholas, and Charles Swanton. 2017. “Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future.” *Cell* 168 (4): 613–28.

- McKenna, Aaron, Gregory M. Findlay, James A. Gagnon, Marshall S. Horwitz, Alexander F. Schier, and Jay Shendure. 2016. "Whole-Organism Lineage Tracing by Combinatorial and Cumulative Genome Editing." *Science* 353 (6298): aaf7907.
- McKenna, Aaron, and James A. Gagnon. 2019. "Recording Development with Single Cell Dynamic Lineage Tracing." *Development* 146 (12). <https://doi.org/10.1242/dev.169730>.
- Min, Jiaqi, Qian Feng, Wenjun Liao, Yiming Liang, Chengwu Gong, Enliang Li, Wenfeng He, Rongfa Yuan, and Linqun Wu. 2018. "IFITM3 Promotes Hepatocellular Carcinoma Invasion and Metastasis by Regulating MMP9 through p38/MAPK Signaling." *FEBS Open Bio* 8 (8): 1299–1311.
- Moore, Luiza, Alex Cagan, Tim H. H. Coorens, Matthew D. C. Neville, Rashesh Sanghvi, Mathijs A. Sanders, Thomas R. W. Oliver, et al. 2021. "The Mutational Landscape of Human Somatic and Germline Cells." *Nature* 597 (7876): 381–86.
- Nam, Anna S., Kyu-Tae Kim, Ronan Chaligne, Franco Izzo, Chelston Ang, Justin Taylor, Robert M. Myers, et al. 2019. "Somatic Mutations and Cell Identity Linked by Genotyping of Transcriptomes." *Nature* 571 (7765): 355–60.
- Navin, Nicholas, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, et al. 2011. "Tumour Evolution Inferred by Single-Cell Sequencing." *Nature* 472 (7341): 90–94.
- Naxerova, Kamila. 2021. "Mutation Fingerprints Encode Cellular Histories." *Nature*.
- Naxerova, Kamila, and Rakesh K. Jain. 2015. "Using Tumour Phylogenetics to Identify the Roots of Metastasis in Humans." *Nature Reviews. Clinical Oncology* 12 (5): 258–72.
- Naxerova, Kamila, Johannes G. Reiter, Elena Brachtel, Jochen K. Lennerz, Marc van de Wetering, Andrew Rowan, Tianxi Cai, et al. 2017. "Origins of Lymphatic and Distant Metastases in Human Colorectal Cancer." *Science* 357 (6346): 55–60.
- "NCI Budget and Appropriations." 2015. April 24, 2015. <https://www.cancer.gov/about-nci/budget>.
- Nieto, M. Angela. 2013. "Epithelial Plasticity: A Common Theme in Embryonic and Cancer Cells." *Science* 342 (6159): 1234850.
- Nieto, M. Angela, Ruby Yun-Ju Huang, Rebecca A. Jackson, and Jean Paul Thiery. 2016. "EMT: 2016." *Cell* 166 (1): 21–45.
- Ocaña, Oscar H., Rebeca Córcoles, Ángels Fabra, Gema Moreno-Bueno, Hervé Acloque, Sonia Vega, Alejandro Barrallo-Gimeno, Amparo Cano, and M. Angela Nieto. 2012. "Metastatic Colonization Requires the Repression of the Epithelial-Mesenchymal Transition Inducer Prrx1." *Cancer Cell*. <https://doi.org/10.1016/j.ccr.2012.10.012>.

- Ostertag, E. M., and H. H. Kazazian Jr. 2001. "Biology of Mammalian L1 Retrotransposons." *Annual Review of Genetics* 35: 501–38.
- Padua, David, Xiang H-F Zhang, Qiongqing Wang, Cristina Nadal, William L. Gerald, Roger R. Gomis, and Joan Massagué. 2008. "TGF $\beta$  Primes Breast Tumors for Lung Metastasis Seeding through Angiopoietin-like 4." *Cell* 133 (1): 66–77.
- Park, Jihye, Jung Min Lim, Inkyung Jung, Seok-Jae Heo, Jinman Park, Yoojin Chang, Hui Kwon Kim, et al. 2021. "Recording of Elapsed Time and Temporal Information about Biological Events Using Cas9." *Cell* 184 (4): 1047–63.e23.
- Park, Seongyeol, Nanda Maya Mali, Ryul Kim, Jeong-Woo Choi, Junehawk Lee, Joonoh Lim, Jung Min Park, et al. 2021. "Clonal Dynamics in Early Human Embryogenesis Inferred from Somatic Mutation." *Nature* 597 (7876): 393–97.
- Pastushenko, levgenia, and Cédric Blanpain. 2019. "EMT Transition States during Tumor Progression and Metastasis." *Trends in Cell Biology* 29 (3): 212–26.
- Pastushenko, levgenia, Audrey Brisebarre, Alejandro Sifrim, Marco Fioramonti, Tatiana Revenco, Soufiane Boumahdi, Alexandra Van Keymeulen, et al. 2018. "Identification of the Tumour Transition States Occurring during EMT." *Nature* 556 (7702): 463–68.
- Pei, Weike, Thorsten B. Feyerabend, Jens Rössler, Xi Wang, Daniel Postrach, Katrin Busch, Immanuel Rode, et al. 2017. "Polylox Barcoding Reveals Haematopoietic Stem Cell Fates Realized in Vivo." *Nature*, August. <https://doi.org/10.1038/nature23653>.
- Pereira, A. A. L., J. F. M. Rego, V. Morris, M. J. Overman, C. Eng, C. R. Garrett, A. T. Boutin, et al. 2015. "Association between KRAS Mutation and Lung Metastasis in Advanced Colorectal Cancer." *British Journal of Cancer* 112 (3): 424–28.
- Perkel, Jeffrey M. 2021. "Single-Cell Analysis Enters the Multiomics Age." *Nature* 595 (7868): 614–16.
- Perli, Samuel D., Cheryl H. Cui, and Timothy K. Lu. 2016. "Continuous Genetic Recording with Self-Targeting CRISPR-Cas in Human Cells." *Science* 353 (6304). <https://doi.org/10.1126/science.aag0511>.
- Popic, Victoria, Raheleh Salari, Iman Hajirasouliha, Dorna Kashef-Haghighi, Robert B. West, and Serafim Batzoglou. 2015. "Fast and Scalable Inference of Multi-Sample Cancer Lineages." *Genome Biology* 16 (May): 91.
- Port, Fillip, and Simon L. Bullock. 2016. "Augmenting CRISPR Applications in Drosophila with tRNA-Flanked sgRNAs." *Nature Methods* 13 (10): 852–54.
- Quinn, Jeffrey J., Matthew G. Jones, Ross A. Okimoto, Shigeki Nanjo, Michelle M. Chan, Nir Yosef, Trevor G. Bivona, and Jonathan S. Weissman. 2021. "Single-Cell

- Lineages Reveal the Rates, Routes, and Drivers of Metastasis in Cancer Xenografts." *Science* 371 (6532). <https://doi.org/10.1126/science.abc1944>.
- Quintana, Elsa, Mark Shackleton, Michael S. Sabel, Douglas R. Fullen, Timothy M. Johnson, and Sean J. Morrison. 2008. "Efficient Tumour Formation by Single Human Melanoma Cells." *Nature* 456 (7222): 593–98.
- Raj, Bushra, James A. Gagnon, and Alexander F. Schier. 2018. "Large-Scale Reconstruction of Cell Lineages Using Single-Cell Readout of Transcriptomes and CRISPR–Cas9 Barcodes by scGESTALT." *Nature Protocols* 541 (October): 331.
- Raj, Bushra, Daniel E. Wagner, Aaron McKenna, Shristi Pandey, Allon M. Klein, Jay Shendure, James A. Gagnon, and Alexander F. Schier. 2018. "Simultaneous Single-Cell Profiling of Lineages and Cell Types in the Vertebrate Brain." *Nature Biotechnology* 40 (March): 181.
- Ramanathan, Arvind, and Stuart L. Schreiber. 2009. "Direct Control of Mitochondrial Function by mTOR." *Proceedings of the National Academy of Sciences of the United States of America* 106 (52): 22229–32.
- "Reflecting on 20 Years of Progress." 2021. *Nature Reviews. Cancer* 21 (10): 605.
- Rhim, Andrew D., Emily T. Mirek, Nicole M. Aiello, Anirban Maitra, Jennifer M. Bailey, Florencia McAllister, Maximilian Reichert, et al. 2012. "EMT and Dissemination Precede Pancreatic Tumor Formation." *Cell* 148 (1-2): 349–61.
- Ritchie, Hannah, and Max Roser. 2019. "Age Structure." *Our World in Data*, September. <https://ourworldindata.org/age-structure>.
- Rodriguez-Fraticelli, Alejo E., Samuel L. Wolock, Caleb S. Weinreb, Riccardo Panero, Sachin H. Patel, Maja Jankovic, Jianlong Sun, Raffaele A. Calogero, Allon M. Klein, and Fernando D. Camargo. 2018. "Clonal Analysis of Lineage Fate in Native Haematopoiesis." *Nature* 553 (7687): 212–16.
- Sancak, Yasemin, Timothy R. Peterson, Yoav D. Shaul, Robert A. Lindquist, Carson C. Thoreen, Liron Bar-Peled, and David M. Sabatini. 2008. "The Rag GTPases Bind Raptor and Mediate Amino Acid Signaling to mTORC1." *Science* 320 (5882): 1496–1501.
- Scheel, Christina, and Robert A. Weinberg. 2012. "Cancer Stem Cells and Epithelial–mesenchymal Transition: Concepts and Molecular Links." *Seminars in Cancer Biology* 22 (5): 396–403.
- Schwab, Annemarie, Aarif Siddiqui, Maria Eleni Vazakidou, Francesca Napoli, Martin Böttcher, Bianca Menchicchi, Umar Raza, et al. 2018. "Polyol Pathway Links Glucose Metabolism to the Aggressiveness of Cancer Cells." *Cancer Research* 78 (7): 1604–18.



- Schwartz, Russell, and Alejandro A. Schäffer. 2017. "The Evolution of Tumour Phylogenetics: Principles and Practice." *Nature Reviews. Genetics* 18 (4): 213–29.
- SEER. 2017. "SEER Cancer Statistics Review 1975-2017." [https://seer.cancer.gov/archive/csr/1975\\_2017/results\\_merged/topic\\_survival\\_by\\_year\\_dx.pdf](https://seer.cancer.gov/archive/csr/1975_2017/results_merged/topic_survival_by_year_dx.pdf).
- Serin Harmanci, Akdes, Arif O. Harmanci, and Xiaobo Zhou. 2020. "CaSpER Identifies and Visualizes CNV Events by Integrative Analysis of Single-Cell or Bulk RNA-Sequencing Data." *Nature Communications* 11 (1): 89.
- Sethi, Nilay, Xudong Dai, Christopher G. Winter, and Yibin Kang. 2011. "Tumor-Derived JAGGED1 Promotes Osteolytic Bone Metastasis of Breast Cancer by Engaging Notch Signaling in Bone Cells." *Cancer Cell* 19 (2): 192–205.
- Shackleton, Mark, Elsa Quintana, Eric R. Fearon, and Sean J. Morrison. 2009. "Heterogeneity in Cancer: Cancer Stem Cells versus Clonal Evolution." *Cell* 138 (5): 822–29.
- Shi, Hubing, Willy Hugo, Xiangju Kong, Aayoung Hong, Richard C. Koya, Gatién Moriceau, Thine Chodon, et al. 2014. "Acquired Resistance and Clonal Evolution in Melanoma during BRAF Inhibitor Therapy." *Cancer Discovery* 4 (1): 80–93.
- Simeonov, Kamen P., China N. Byrns, Megan L. Clark, Robert J. Norgard, Beth Martin, Ben Z. Stanger, Jay Shendure, Aaron McKenna, and Christopher J. Lengner. 2021. "Single-Cell Lineage Tracing of Metastatic Cancer Reveals Selection of Hybrid EMT States." *Cancer Cell* 39 (August): 1–13.
- Sizemore, Gina M., Jason R. Pitarresi, Subhasree Balakrishnan, and Michael C. Ostrowski. 2017. "The ETS Family of Oncogenic Transcription Factors in Solid Tumours." *Nature Reviews. Cancer* 17 (6): 337–51.
- Snippert, Hugo J., Laurens G. van der Flier, Toshiro Sato, Johan H. van Es, Maaïke van den Born, Carla Kroon-Veenboer, Nick Barker, et al. 2010. "Intestinal Crypt Homeostasis Results from Neutral Competition between Symmetrically Dividing Lgr5 Stem Cells." *Cell* 143 (1): 134–44.
- Spanjaard, Bastiaan, Bo Hu, Nina Mitic, Pedro Olivares-Chauvet, Sharan Janjuha, Nikolay Ninov, and Jan Philipp Junker. 2018. "Simultaneous Lineage Tracing and Cell-Type Identification Using CRISPR-Cas9-Induced Genetic Scars." *Nature Biotechnology*, April. <https://doi.org/10.1038/nbt.4124>.
- Stemmler, Marc P., Rebecca L. Eccles, Simone Brabletz, and Thomas Brabletz. 2019. "Non-Redundant Functions of EMT Transcription Factors." *Nature Cell Biology* 21 (1): 102–12.
- Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal. 2009. "The Cancer Genome." *Nature* 458 (7239): 719–24.

- Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck 3rd, Yuhao Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. "Comprehensive Integration of Single-Cell Data." *Cell* 177 (7): 1888–1902.e21.
- Sun, Jianlong, Azucena Ramos, Brad Chapman, Jonathan B. Johnnidis, Linda Le, Yu-Jui Ho, Allon Klein, Oliver Hofmann, and Fernando D. Camargo. 2014. "Clonal Dynamics of Native Haematopoiesis." *Nature* 514 (7522): 322–27.
- Takano, Shigetsugu, Maximilian Reichert, Basil Bakir, Koushik K. Das, Takahiro Nishida, Masaru Miyazaki, Steffen Heeg, et al. 2016. "Prrx1 Isoform Switching Regulates Pancreatic Cancer Invasion and Metastatic Colonization." *Genes & Development* 30 (2): 233–47.
- Takebe, Naoko, Pamela J. Harris, Ronald Q. Warren, and S. Percy Ivy. 2011. "Targeting Cancer Stem Cells by Inhibiting Wnt, Notch, and Hedgehog Pathways." *Nature Reviews. Clinical Oncology* 8 (2): 97–106.
- Talmadge, J. E., S. R. Wolman, and I. J. Fidler. 1982. "Evidence for the Clonal Origin of Spontaneous Metastases." *Science* 217 (4557): 361–63.
- Thomson, Timothy M., Cristina Balcells, and Marta Cascante. 2019. "Metabolic Plasticity and Epithelial-Mesenchymal Transition." *Journal of Clinical Medicine Research* 8 (7). <https://doi.org/10.3390/jcm8070967>.
- Tian, Chenxi, Karl R. Clauser, Daniel Öhlund, Steffen Rickelt, Ying Huang, Mala Gupta, D. R. Mani, Steven A. Carr, David A. Tuveson, and Richard O. Hynes. 2019. "Proteomic Analyses of ECM during Pancreatic Ductal Adenocarcinoma Progression Reveal Different Contributions by Tumor and Stromal Cells." *Proceedings of the National Academy of Sciences of the United States of America* 116 (39): 19609–18.
- Todaro, Matilde, Mileidys Perez Alea, Anna B. Di Stefano, Patrizia Cammareri, Louis Vermeulen, Flora Iovino, Claudio Tripodo, et al. 2007. "Colon Cancer Stem Cells Dictate Tumor Growth and Resist Cell Death by Production of Interleukin-4." *Cell Stem Cell* 1 (4): 389–402.
- Tsai, Jeff H., Joana Liu Donaher, Danielle A. Murphy, Sandra Chau, and Jing Yang. 2012. "Spatiotemporal Regulation of Epithelial-Mesenchymal Transition Is Essential for Squamous Cell Carcinoma Metastasis." *Cancer Cell* 22 (6): 725–36.
- Turajlic, Samra, and Charles Swanton. 2016. "Metastasis as an Evolutionary Process." *Science* 352 (6282): 169–75.
- Uchi, Ryutaro, Yusuke Takahashi, Atsushi Niida, Teppei Shimamura, Hidenari Hirata, Keishi Sugimachi, Genta Sawada, et al. 2016. "Integrated Multiregional Analysis Proposing a New Model of Colorectal Cancer Evolution." *PLoS Genetics* 12 (2): e1005778.

- Wang, Hao, Hong-Sheng Wang, Bin-Hua Zhou, Cui-Lin Li, Fan Zhang, Xian-Feng Wang, Ge Zhang, Xian-Zhang Bu, Shao-Hui Cai, and Jun Du. 2013. "Epithelial-Mesenchymal Transition (EMT) Induced by TNF- $\alpha$  Requires AKT/GSK-3 $\beta$ -Mediated Stabilization of Snail in Colorectal Cancer." *PLoS One* 8 (2): e56664.
- Weinreb, Caleb, Alejo Rodriguez-Fraticelli, Fernando D. Camargo, and Allon M. Klein. 2020. "Lineage Tracing on Transcriptional Landscapes Links State to Fate during Differentiation." *Science*, January. <https://doi.org/10.1126/science.aaw3381>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemond, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686.
- Wu, C. X., A. Xu, C. C. Zhang, P. Olson, and L. Chen. 2017. "Notch Inhibitor PF-03084014 Inhibits Hepatocellular Carcinoma Growth and Metastasis via Suppression of Cancer Stemness due to Reduced Activation of Notch1–Stat3." *Molecular Cancer*. <https://mct.aacrjournals.org/content/16/8/1531.short>.
- Yachida, Shinichi, Siân Jones, Ivana Bozic, Tibor Antal, Rebecca Leary, Baojin Fu, Mihoko Kamiyama, et al. 2010. "Distant Metastasis Occurs Late during the Genetic Evolution of Pancreatic Cancer." *Nature* 467 (7319): 1114–17.
- Yang, Jing, Parker Antin, Geert Berx, Cédric Blanpain, Thomas Brabletz, Marianne Bronner, Kyra Campbell, et al. 2020. "Guidelines and Definitions for Research on Epithelial–mesenchymal Transition." *Nature Reviews. Molecular Cell Biology* 21 (6): 341–52.
- Yang, Jing, Sendurai A. Mani, Joana Liu Donaher, Sridhar Ramaswamy, Raphael A. Itzykson, Christophe Come, Pierre Savagner, Inna Gitelman, Andrea Richardson, and Robert A. Weinberg. 2004. "Twist, a Master Regulator of Morphogenesis, Plays an Essential Role in Tumor Metastasis." *Cell* 117 (7): 927–39.
- Yauch, Robert L., Stephen E. Gould, Suzie J. Scales, Tracy Tang, Hua Tian, Christina P. Ahn, Derek Marshall, et al. 2008. "A Paracrine Requirement for Hedgehog Signalling in Cancer." *Nature* 455 (7211): 406–10.
- Yu, Fang, Dan Xie, Samuel S. Ng, Ching Tung Lum, Mu-Yan Cai, William K. Cheung, Hsiang-Fu Kung, Guimiao Lin, Xiaomei Wang, and Marie C. Lin. 2015. "IFITM1 Promotes the Metastasis of Human Colorectal Cancer via CAV-1." *Cancer Letters* 368 (1): 135–43.
- Zavadil, Jiri, and Erwin P. Böttinger. 2005. "TGF-Beta and Epithelial-to-Mesenchymal Transitions." *Oncogene* 24 (37): 5764–74.
- Zhang, Jingyu, Xiao-Jun Tian, and Jianhua Xing. 2016. "Signal Transduction Pathways of EMT Induced by TGF- $\beta$ , SHH, and WNT and Their Crosstalks." *Journal of Clinical Medicine Research* 5 (4). <https://doi.org/10.3390/jcm5040041>.

- Zhang, Jingyu, Xiao-Jun Tian, Hang Zhang, Yue Teng, Ruoyan Li, Fan Bai, Subbiah Elankumaran, and Jianhua Xing. 2014. "TGF- $\beta$ -Induced Epithelial-to-Mesenchymal Transition Proceeds through Stepwise Activation of Multiple Feedback Loops." *Science Signaling* 7 (345): ra91.
- Zhang, Weijie, Igor L. Bado, Jingyuan Hu, Ying-Wooi Wan, Ling Wu, Hai Wang, Yang Gao, et al. 2021. "The Bone Microenvironment Invigorates Metastatic Seeds for Further Dissemination." *Cell* 184 (9): 2471–86.e20.
- Zheng, Xiaofeng, Julienne L. Carstens, Jiha Kim, Matthew Scheible, Judith Kaye, Hikaru Sugimoto, Chia-Chin Wu, Valerie S. LeBleu, and Raghuram Kalluri. 2015. "Epithelial-to-Mesenchymal Transition Is Dispensable for Metastasis but Induces Chemoresistance in Pancreatic Cancer." *Nature* 527 (7579): 525–30.
- Zhuang, Xueqian, Hao Zhang, Xiaoyan Li, Xiaoxun Li, Min Cong, Fangli Peng, Jingyi Yu, Xue Zhang, Qifeng Yang, and Guohong Hu. 2017. "Differential Effects on Lung and Bone Metastasis of Breast Cancer by Wnt Signaling Inhibitor DKK1." *Nature Cell Biology* 19 (10): 1274–85.