nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0489-5

A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition

José L. McFaline-Figueroa[1], Andrew J. Hill[1], Xiaojie Qiu[1,2], Dana Jackson[1], Jay Shendure[1,3,4,5] and Cole Trapnell[1,3,5*]

Integrating single-cell trajectory analysis with pooled genetic screening could reveal the genetic architecture that guides cellular decisions in development and disease. We applied this paradigm to probe the genetic circuitry that controls epithelial-to-mesenchymal transition (EMT). We used single-cell RNA sequencing to profile epithelial cells undergoing a spontaneous spatially determined EMT in the presence or absence of transforming growth factor-β. Pseudospatial trajectory analysis identified continuous waves of gene regulation as opposed to discrete 'partial' stages of EMT. KRAS was connected to the exit from the epithelial state and the acquisition of a fully mesenchymal phenotype. A pooled single-cell CRISPR-Cas9 screen identified EMT-associated receptors and transcription factors, including regulators of KRAS, whose loss impeded progress along the EMT. Inhibiting the KRAS effector MEK and its upstream activators EGFR and MET demonstrates that interruption of key signaling events reveals regulatory 'checkpoints' in the EMT continuum that mimic discrete stages, and reconciles opposing views of the program that controls EMT.

During EMT, cells dissolve strong contacts and leave organized sheets, shifting from apical–basal to front–rear polarity. As they become mesenchymal, their motility and ability to break down extracellular matrix enables them to invade surrounding tissue[1,2]. EMT is fundamental to development[3], wound healing[4,5] and the metastatic dissemination of tumor cells[2,5,6].

Several studies have identified discrete intermediate 'stages' of EMT based on the expression of a handful of marker genes[7–9]. However, recent single-cell mass cytometry and RNA-seq analyses of breast cancer cells suggest that they fall along a continuum[10,11]. As such, it remains unclear whether or not cells exist in functionally discrete states during EMT, and the genetic circuitry that controls the transition remains incompletely defined. Partial EMT is implicated in renal fibrosis[12,13] and pancreatic ductal adenocarcinoma[14] and is positively correlated with tumor grade and metastatic potential in head and neck squamous cell carcinoma (HNSCC)[15]. Characterizing the nature of intermediate EMT would have an immediate impact on our understanding of disease.

Here, we apply single-cell RNA sequencing (scRNA-seq) coupled with unsupervised machine learning techniques[16,17] to analyze a 'pseudospatial'[18] model of EMT and identify regulators of its progression. We analyze a two-dimensional (2D) model system of spontaneous confluence-dependent EMT in human mammary epithelial cells[19]. Cells fell continuously along a gradient of EMT progression, revealing distinct waves of gene regulation. We incorporate a pooled single-cell CRISPR-Cas9 screen into our pseudospatial trajectory analysis to determine the dependency of EMT-associated signaling events on progression along the EMT continuum. These experiments uncover a hierarchy of transcription factors and cell surface receptors that drive cells through EMT. Loss-of-function of one of several surface receptors slows the progress through EMT, explaining how cells transiting through a continuous process appear to be in one of several discrete stages in some experimental systems.

Results

Pseudospatial trajectory analysis of spontaneous EMT. To define the transcriptional program executed by normal human cells undergoing EMT, we devised an in vitro system in which cells from an epithelial colony migrate into unoccupied margins of the plate (Fig. 1a). We seeded MCF10A mammary epithelial cells[19] within cloning rings as a high-confluence patch in the center of a tissue culture dish. We then removed the rings after which cells at the border can sense adjacent unoccupied space and spontaneously undergo an EMT. The spontaneous EMT in this system is analogous to that observed for MCF10A cells on wounding in scratch-wound healing assays[20,21]. Cells at the periphery of the patch acquired a spindle-like morphology and formed leading and protruding edges consistent with the acquisition of a mesenchymal phenotype (Supplementary Fig. 1). Cells collected from a single well of our assay expressed levels of E-cadherin and vimentin protein spanning a dynamic range that included those cultured at low or high confluence (Supplementary Fig. 1c,d). We dissected the patch to isolate 'inner' cells (2,440 cells) and 'outer' cells (2,564 cells). Inner and outer fractions were dissociated into single-cell suspensions and subjected to scRNA-seq on the 10x Chromium platform (Fig. 1a and Supplementary Table 1).

Unsupervised clustering with *t*-distributed stochastic neighbor embedding (t-SNE) separated cells from inner and outer fractions (Fig. 1b), and expression of the mesenchymal marker *VIM* varied in a reciprocal gradient to the epithelial markers *CDH1* and *DSP* across embedded cells (Fig. 1c and Supplementary Fig. 2). However, we did not observe separated clusters of cells along this

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [2]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, CA, USA. [3]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA. [4]Howard Hughes Medical Institute, Seattle, WA, USA. [5]Allen Discovery Center for Cell Lineage Tracing, Seattle, WA, USA. *e-mail: coletrap@uw.edu

axis of epithelial and mesenchymal marker expression, suggesting continual progression along an EMT rather than a sequence of discrete stages.

Individual cells at similar radii from the center of a patch could be in different stages of EMT, analogous to how cells proceed asynchronously through temporal processes such as differentiation. To resolve cellular heterogeneity and recover the program that characterizes the progress of a cell through EMT, we ordered cells using Monocle[16,17]. Monocle organized cells along a linear pseudospatial[18,22] trajectory, with cells from inner and outer fractions concentrated at the beginning and end of its axis, respectively (Fig. 1d and Supplementary Fig. 3). Simulated sampling from the ends of the continuum and repeating our analysis excluded the possibility that this continuity was an artifact of trajectory analysis (Supplementary Fig. 4).

Classic markers of EMT varied in expression over the trajectory. Protein and messenger RNA levels of the epithelial marker E-cadherin (*CDH1*) decreased as cells radiated from the center of the colony and over the pseudospatial trajectory, consistent with a spontaneous spatially determined EMT (Fig. 1e,f). Conversely, mRNA levels of *VIM* increased sharply in cells in the second half of the trajectory (Fig. 1f). Partial or intermediate EMT has classically been defined as the coexpression of epithelial and mesenchymal traits[23,24]. Accordingly, cells positive for both *CDH1* and *VIM* mRNA were most frequent in the second half of the trajectory (Supplementary Fig. 5). The population-level average expression of two epithelial markers, *CDH1* and *CRB3*, did not vary drastically between inner and outer fractions (Supplementary Fig. 6), highlighting the value of single-cell techniques to capture the dynamics of gene regulatory changes associated with EMT.

We next identified genes regulated during EMT by performing differential expression analysis parameterized by the position of each cell along the trajectory (Supplementary Table 2). Clustering the 1,105 differentially expressed genes (DEGs) (likelihood ratio test; false discovery rate (FDR) $< 1 \times 10^{-10}$; area under the curve (AUC) >10 in at least one quantile, see Methods and Supplementary Table 3) revealed six groups of genes with similar kinetics. We performed geneset analysis using the Gene Ontology biological processes[25,26] and MSigDB hallmarks molecular signature[27] geneset collections. Genes in cluster 6 were upregulated and enriched for roles in translational regulation or EMT, while those in downregulated cluster 1 were linked to epidermis development. Cluster 5, highly expressed in the outermost regions of the pseudospatial trajectory, was associated with the regulation of the cell cycle, consistent with the relief of contact-mediated inhibition of proliferation (Fig. 1g,h and Supplementary Table 4).

Geneset analysis identified pathways upstream of pseudospace-dependent gene expression. Cluster 1 was enriched for genes repressed by active KRAS signaling[28,29], including some with roles in EMT. For example, keratin 1 (*KRT1*) was expressed in cells at the epithelial end of the trajectory but silenced as cells approached the border of the patch (Supplementary Fig. 7). Keratins traffic E-cadherin to the cell membrane, while vimentin does not[30], and the shift in cytoskeletal filament composition from keratin- to vimentin-containing is integral to EMT[31]. The EMT-associated cluster 6 included the unfolded protein response (UPR) transcriptional regulator ATF4 whose increased expression preceded upregulation of genes in cluster 5, which was enriched for genes associated with the UPR (cluster 5 and 6, Fig. 1g,h and Supplementary Fig. 8), consistent with a recent study demonstrating that the induction of EMT elicits protective activation of the UPR[32].

Repeating our spatial EMT assay and single-cell transcriptional profiling using primary human mammary epithelial cells (HuMEC) identified a similar linear pseudospatial trajectory and distribution of inner and outer cells (Supplementary Fig. 9a,b and Supplementary Table 5). The dynamics of epithelial and mesenchymal marker expression was comparable albeit with decreased *CDH1* downregulation and more drastic upregulation of *FN1* (Fig. 1i and Supplementary Fig. 9c). Having identified a spatial EMT in another epithelial cell type we sought to understand how this phenotype changes in response to a strong inducer of EMT.

Pseudospatial trajectory alignment elucidates transforming growth factor β (TGF-β)-driven full EMT. Activation of the TGF-β pathway leads to a powerful induction of EMT[33,34]. We repeated our pseudospace experiment, this time treating cells with TGF-β to promote mesenchymal conversion in MCF10A cells[7]. We sequenced transcriptomes of 2,121 inner and 2,116 outer colony cells that were segregated in t-SNE space but did not form coherent clusters, and whose expressed *FN1* and *VIM* continuously varied (Supplementary Fig. 10). Thus, adding a strong extracellular signal promoting EMT did not drive cells into discrete stages. We therefore constructed a pseudospatial trajectory for TGF-β as well (Supplementary Fig. 11).

To compare cells from spontaneous and TGF-β-driven EMT trajectories, we used trajectory alignment[35–37], a technique that employs Dynamic Time Warping[38,39] to map cells onto a common pseudospatial axis (Fig. 2a). Along the aligned axis, *CDH1* and *CRB3* were expressed in cells treated with TGF-β with similar kinetics to those undergoing confluence-mediated EMT (Fig. 2b), and consistent with reports that maintenance of cell–cell contacts prevents TGF-β stimulation from fully repressing an epithelial phenotype[40]. However, TGF-β exposure is sufficient to drive the expression of mesenchymal genes even in cells within the epithelial core. Additionally, only cells treated with TGF-β and positioned at the outer extreme of the trajectory expressed robust levels of *FN1* and *CDH2*, suggesting a full E- to N-cadherin switch. Exposure of HuMEC cells to TGF-β similarly led to a robust increase in *VIM* and *FN1* at the beginning of the trajectory (Supplementary Fig. 12c); however, expression of *CDH2* was not apparent. A broader geneset analysis comparing normalized average expression scores[41] showed that TGF-β drove MSigDB Hallmark EMT genes higher and Gene Ontology biological process epidermis development genes lower in both MCF10A and HuMEC cells (Supplementary Fig. 13).

To identify genes responsive to TGF-β, we tested for differential expression as a function of TGF-β treatment, subtracting changes attributable to pseudospatial position. This analysis identified 1,328 genes in 10 clusters with distinct TGF-β-dependent dynamics (Fig. 2c, likelihood ratio test; FDR $< 1 \times 10^{-10}$ and $|\Delta AUC| > 0.02$, see Methods, Supplementary Fig. 14 and Supplementary Table 6). For example, cluster 5 contained cell-cycle-related genes upregulated along both trajectories (Fig. 2c and Supplementary Table 7). Cluster 4 contained genes upregulated toward the end of the spontaneous trajectory but maintained at high levels throughout the TGF-β-mediated trajectory (Fig. 2c). This cluster included two EMT-associated genes, one of which, *NNMT*, is a marker of the metabolic changes that accompany EMT[42] (Fig. 2d). In contrast, clusters 6 and 8 contained EMT genes that peaked at the middle or end of the TGF-β-driven trajectory, respectively (Fig. 2c,e,f), but were unaltered or induced to a lesser degree in the spontaneous trajectory. Therefore, cells at comparable positions in spontaneous versus TGF-β-mediated EMT continua as defined by epithelial markers display pronounced transcriptional differences.

To explore which molecular regulators are responsible for shared and distinct patterns of spontaneous and TGF-β-mediated gene regulation during EMT we performed geneset analysis using the MSigDB Oncogenic Signature geneset collection. This geneset collection is composed of genes whose expression increases or decreases as a function of perturbing signaling pathways[43]. Cluster 8 included genes upregulated as cells treated with TGF-β undergo EMT but are weakly altered during spontaneous EMT. These were enriched for genes expressed in response to KRAS signaling[28], including genes with roles in EMT, such as *CXCL1* and *CXCL2*, which induce

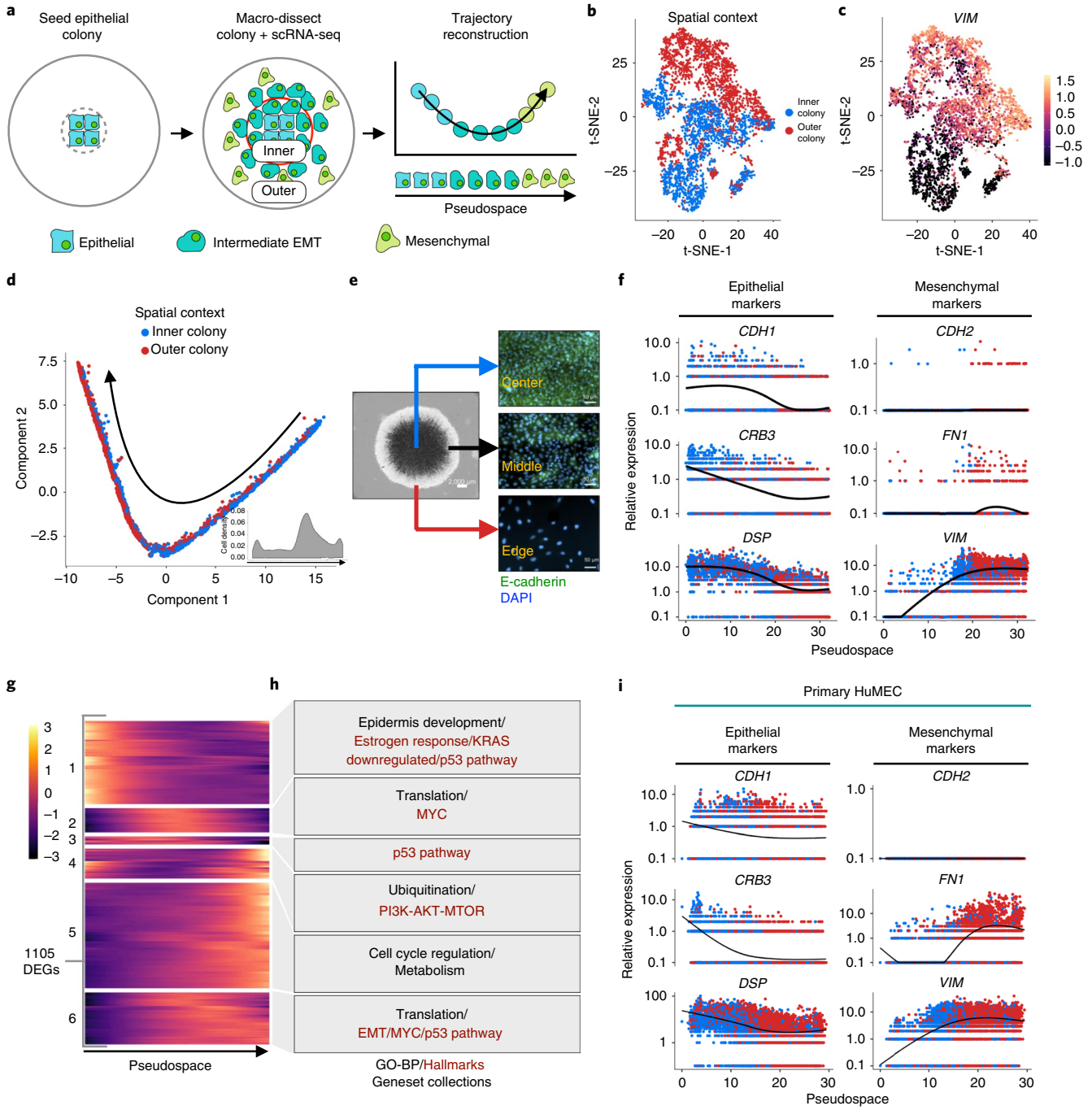


Fig. 1 | Pseudospatial trajectory reconstruction of spontaneous EMT reveals the transition as a continuum of epithelial–mesenchymal states. **a**, Schematic of spontaneous confluence-dependent EMT assay, cell isolation and pseudospatial trajectory reconstruction using Monocle2. Red circle denotes the area that defines inner and outer cells for macro-dissection. **b**,**c**, t-SNE of cells from our spontaneous EMT assay. Cells are colored according to the fraction from which they were isolated (**b**) or expression of the mesenchymal marker *VIM* (**c**). **d**, Trajectory of inner and outer MCF10A cells on spontaneous EMT progression. Arrow denotes progression of pseudospace. Insert, density of cells across pseudospace. **e**, Left, stitched brightfield images of an MCF10A colony at the end of our spontaneous EMT assay (scale bar, 2,000 µm). Right, top to bottom, E-cadherin and DAPI staining of cells from the center, middle and edge of the MCF10A colony (scale bar, 50 µm; representative fields from six images across three independent samples). **f**, Expression of epithelial and mesenchymal markers across pseudospace. Cells are colored as in **b**. **g**, Hierarchical clustering of kinetic curves for dynamically regulated genes across pseudospace for all 5,004 cells in our experiment (likelihood ratio test, FDR < 1×10^{-10} and AUC > 10). Rows represent row centered dynamics of gene expression. **h**, Geneset analysis using the Gene Ontology Biological Processes (GO-BP) and MSigDB Hallmarks geneset collections of gene clusters from **g** (hypergeometric test, FDR < 0.05). **i**, Expression of epithelial and mesenchymal markers across pseudospace in primary human mammary epithelial cells (HuMEC). Cells are colored as in **b**.

cellular migration[44,45] (Supplementary Fig. 15). Conversely, cluster 10 included epithelial marker genes downregulated early in spontaneous EMT and expressed at low levels in cells treated with TGF-β (for example *KRT4* and *KRT16*) (Supplementary Fig. 16). These and several others are known to be repressed by active KRAS signaling[28]. This observation, together with pathway analysis of spontaneous

EMT, implies that KRAS signaling is sustained throughout both spontaneous and TGF-β-driven transitions, suggesting it governs multiple points of the EMT continuum in normal cells.

Single-cell flow cytometric profiling of TGF-β-induced EMT described the transition as a three-state process[7,46]. In contrast to this 'discrete' view, we observed a continuous trajectory over which cells were distributed and along which many genes, including classic markers of epithelial and mesenchymal states, exhibit smooth changes in expression. Few cells undergoing spontaneous EMT expressed high levels of some mesenchymal markers, raising the possibility that we failed to capture some discrete, physiologically important 'stages' of EMT. However, exposing cells to TGF-β also drives them over a continuum, albeit one with different spatial patterns of transcriptional regulation.

To investigate whether tumor cells in vivo transit through an EMT continuum similar to the one observed in vitro, we re-analyzed scRNA-seq data from patients with HNSCC[15]. The most mesenchymal tumor, as ranked by Puram et al[15]., expressed EMT genes at similar levels to cells at the outer end of our TGF-β-driven pseudospatial trajectory (Supplementary Fig. 17). Genes that make up early and late waves of KRAS-associated EMT in vitro (cluster 10 and 8, respectively, Fig. 2c) were expressed in a manner consistent with their partial EMT phenotypes assigned by Puram et al. (Fig. 2g,h). To confirm that the similarity between our in vitro model and the tumor cells was not limited to known EMT genes, we projected tumor cells onto our spontaneous and TGF-β-driven trajectories based on full transcriptome signatures using a nearest-neighbor matching algorithm[47] (Methods). Most tumor cells mapped to the end of our spontaneous EMT trajectory. In contrast, tumor cells projected more uniformly over the TGF-β-driven trajectory (Fig. 2i). Individual tumors covered a substantial range of the trajectory, suggesting our TGF-β-driven model captures much of the transcriptional diversity present in a single patient sample. Finally, we tested whether Monocle2 could reconstruct pseudospatial trajectories directly from tumor cells. For three of the four tumors with sufficient cells for Monocle analysis, the algorithm recovered a linear trajectory (Supplementary Fig. 18) with similar expression kinetics to in vitro trajectories (Supplementary Fig. 18). Taken together, these analyses suggest that the waves of gene regulation that occur during EMT in vitro occur to varying extents in vivo.

A pooled loss-of-function screen identifies genes regulating EMT progression. We reasoned that certain regulators control passage through parts of the EMT continuum and a lack of one or more of these signals leads to accumulation of cells at 'discrete' EMT 'stages'. To identify regulators of progression along the continuum, we devised a high-throughput loss-of-function screen to probe the architecture of pathways with known involvement in EMT. Several groups recently devised methods for coupling CRISPR-based screens and a scRNA-seq readout, thereby capturing the identity of the single-guide RNA(s) (sgRNAs) that a cell received in conjunction with its gene expression profile[48–52]. Here we used a modified version of CRISPR droplet sequencing (CROP-seq)[52], which does not rely on the pairing of sgRNAs with distally located barcodes. We recently showed that this design is preferable to alternatives, avoiding template switching between sgRNAs and associated barcodes during lentiviral co-packaging[53].

We selected 16 cell surface receptors and 24 transcription factors for screening via CROP-seq in our 2D EMT system (Fig. 3a). These targets include receptors reported to activate KRAS (*EGFR*, *MET*, *FGFR1*, *FGFR2*, *ITGAV*, *ITGB1* and *ITGB3*)[54–57] along with others that drive Wnt, Notch and other pathways (Fig. 3b). Transcription factors that activate or repress EMT genes included both well-characterized (*SNAI1/2*, *TWIST1/2* and *ZEB1/2*) and recently reported (*FOXD3*, *GATA6* and *SOX9*) regulators[1]. We repeated our in vitro EMT assay with a mixture of cells edited with sgRNAs to one of the 40 genes (or non-targeting controls, NTC) and subjected them to scRNA-seq after being cultured with TGF-β (12,337 cells) or without (17,093 cells). Unsupervised clustering analysis of cells treated with TGF-β identified prominent, clearly demarcated clusters of cells that retained expression of the epithelial markers *CDH1* and *CRB3* and failed to upregulate *FN1* and *VIM* (Fig. 3c and Supplementary Fig. 19a,b). Cells expressing sgRNAs targeting *TGFBR1* and *TGFBR2* were enriched in these clusters (Fisher's exact test; FDR $< 1 \times 10^{-50}$) (Fig. 3d,e and Supplementary Fig. 19c), while NTC sgRNAs were largely absent from them. Importantly, this distribution was not caused by the number of *TGFBR1* and *TGFBR2* sgRNA cells in our screen (Fig. 3f). Cells with sgRNAs against *TGFBR1* and *TGFBR2* expressed lower levels of *FN1* and *VIM* than those with NTC sgRNAs, indicating a failure to activate a TGF-β-driven EMT (Fig. 3g) and confirming that CROP-seq can be used to identify molecular phenotypes along the EMT continuum.

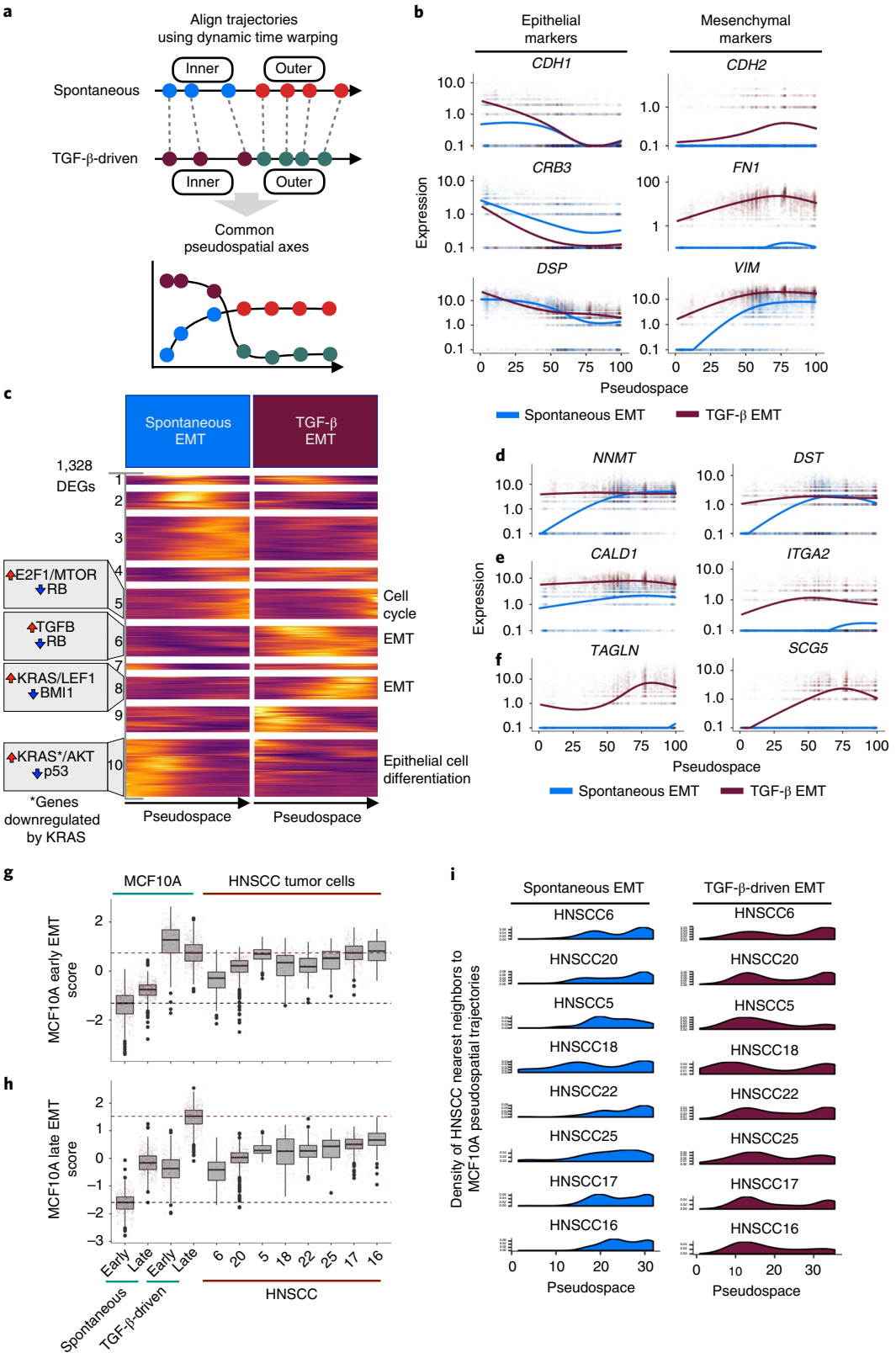
We next sought to organize edited cells into a pseudospatial trajectory. We compared NTC cells from inner and outer fractions, which revealed 1,197 and 761 DEGs in the spontaneous and TGF-β-driven EMT, respectively, more than 80% of which were also found in unedited EMT experiments (Supplementary Fig. 20a–d). Pseudospatial trajectories reconstructed from NTC cells aligned to unedited trajectories with only minimal warping (Supplementary Fig. 20e–g). We then provided Monocle2 with all edited cells, which constructed trajectories along which EMT marker genes were expressed with kinetics similar to unedited cells (Supplementary Fig. 21). Differential expression analysis identified 978 and 4,079 genes that varied across genotypes along spontaneous and TGF-β-driven trajectories, respectively (Supplementary Fig. 22 and Supplementary Tables 8 and 9).

We hypothesized that loss of surface receptors that transduce signals important for EMT, or the transcription factors they drive, would alter a cell's progress along the trajectories. To determine whether loss-of-function of EMT-associated targets altered their progression along pseudospace, we divided the trajectory into bins according to the density of cells along spontaneous and

Fig. 2 | Alignment of spontaneous and TGF-β-driven EMT pseudospatial trajectories identifies discrete waves along the EMT continuum. **a**, Dynamic time warping of pseudospatial trajectories allows for comparison of the dynamics of EMT progression along a common axis. **b**, Epithelial and mesenchymal marker expression across warped pseudospace (cells are colored-coded by treatment). **c**, Hierarchical clustering of kinetic curves for dynamically regulated genes that vary significantly between spontaneous (5,004 cells) and TGF-β-driven (4,237 cells) EMT trajectories (likelihood ratio test, FDR $< 1 \times 10^{-10}$ and $|\Delta AUC| > 0.02$). Rows represent row centered dynamics of gene expression. Left, geneset analysis on gene clusters using the Oncogenic Signatures geneset collection (hypergeometric test, FDR < 0.05). Red and blue arrows denote association with increased or decreased activity, respectively. At right: geneset analysis on gene clusters using the Gene Ontology biological process and Hallmarks geneset collections (hypergeometric test, FDR < 0.05). **d–f**, Pseudospatial expression dynamics of EMT-associated genes that increase in expression at the end of the spontaneous trajectory and are highly expressed across the TGF-β-driven trajectory (**d**), toward the middle of the TGF-β-driven trajectory (**e**) and toward the end of the TGF-β-driven trajectory (**f**). **g–h**, Box plots of early and late EMT scores of MCF10A cells at early and late positions in pseudospatial trajectories (mock, 1,020 cells; TGF-β, 772 cells) and HNSCC tumors (6, 80 cells; 20, 321 cells; 5,41 cells; 18, 140 cells; 22, 119 cells; 25, 54 cells; 17, 330 cells; 16, 56 cells). Box plots depict the median score (bold line within box) with lower and upper hinges depicting the 25th and 75th percentiles, respectively. **i**, Density of cells across EMT trajectories after *k*-nearest-neighbor projection of HNSCC tumor cells to MCF10A cells under spontaneous and TGF-β-driven conditions.

TGF-β-driven EMT trajectories resulting in 7 and 8 bins, respectively. We then tested whether cells carrying sgRNAs against each target altered their distribution over these 'regions' of the aligned trajectories, relative to NTCs. We determined empirical false-discovery rates of these tests by comparing enrichments of knockout cells to a random sampling of NTC cells (Supplementary Fig. 23, see Methods for details).

Of the 40 genes tested, 30 significantly shifted the pseudospatial positions of the cells when targeted via CROP-seq, with 11 overlapping between conditions (Fig. 4a,b and Supplementary Fig. 23).



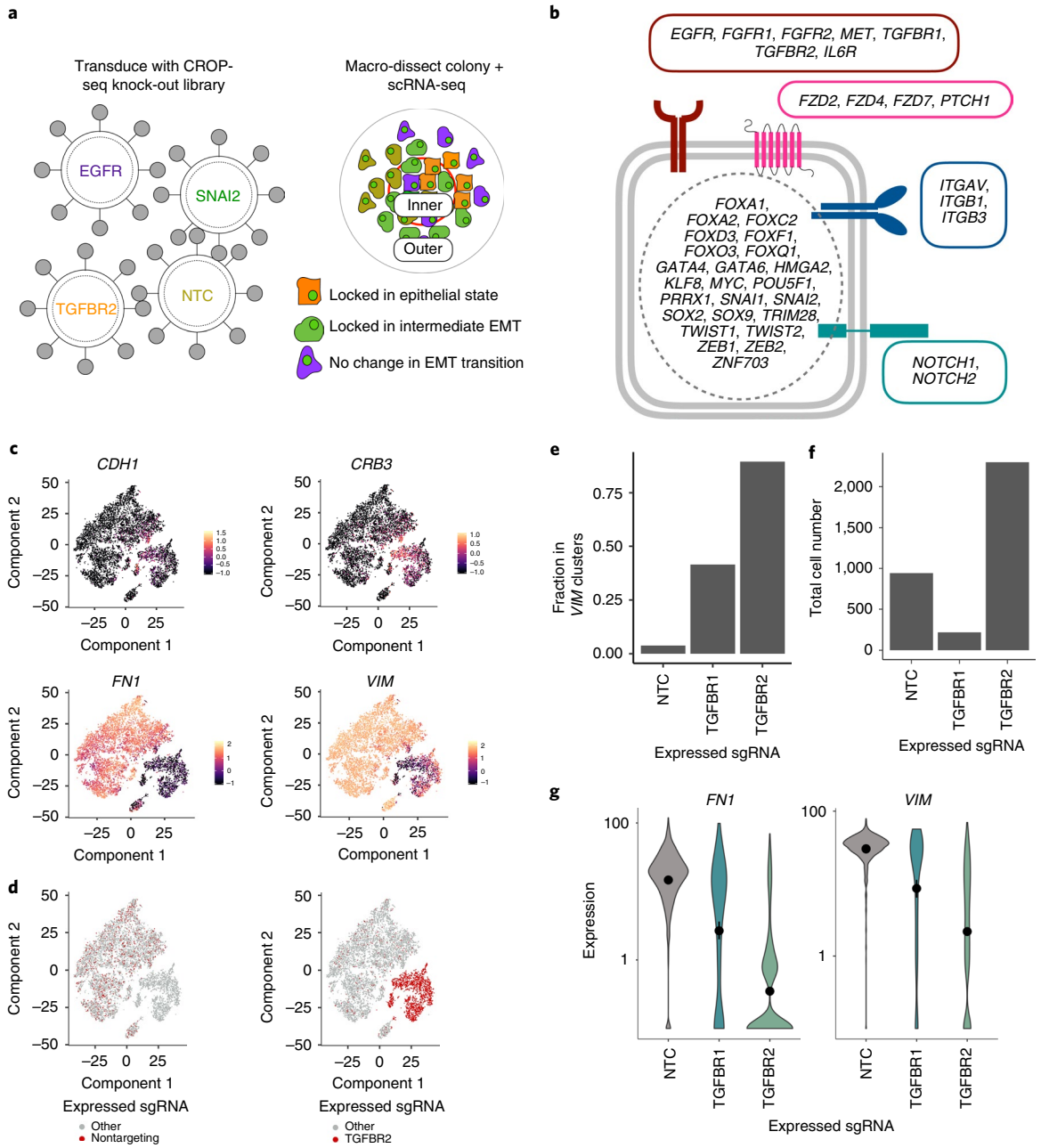


Fig. 3 | Multiplexed loss-of-function screening of EMT-associated genes recovers deficiencies in TGF-β-induced EMT. a, Schematic of pooled approach to determine regulators of distinct EMT states. Red circle in right panel denotes the area that defines the boundary between inner and outer cells for macro-dissection. **b**, Collection of EMT-associated cell surface receptors and transcription factors included in our CROP-seq screen. **c**,**d**, t-SNE of sgRNA containing cells from our TGF-β-exposed CROP-seq experiment colored by EMT marker expression (**c**) or expression of *NTC* or *TGFBR2* sgRNAs (**d**). **e**, Fraction of cells within *VIM* low clusters expressing *NTC*, *TGFBR1* or *TGFBR2* sgRNAs from our TGF-β-exposed CROP-seq screen. **f**, Total number of cells expressing *NTC*, *TGFBR1* or *TGFBR2* sgRNAs from our TGF-β-exposed CROP-seq screen. **g**, Expression of *FN1* and *VIM* across cells expressing a sgRNA to *NTC* (943 cells), *TGFBR1* (219 cells) or *TGFBR2* (2,299 cells) from our TGF-β-exposed CROP-seq experiment. The points within the violin depict the mean expression level for each group with the violin spanning the minimum and maximum expression value across cells.

Some targets were modestly enriched (less than two-fold) at a given pseudospatial position. For example, in the spontaneous EMT trajectory, cells with sgRNAs targeting *FZD7* were enriched at region 1, near the epithelial core of the trajectory, and region 3 (Fig. 4a). Other gene knockouts induced strong, focal accumulation of cells at one or two positions along the EMT continuum (Fig. 4a,b). Loss of EGFR induced focal accumulation at region 3 (Supplementary Fig. 24a). Similarly, cells with sgRNAs against *MET* were enriched in regions 2 and more strongly in region 3. The majority of significantly enriched targets accumulated in region 3 directly preceding a decrease in the total number of *CDH1* single-positive cells and an increase in *CDH1*/*VIM* double-positive cells (Supplementary Fig. 25).

Edited cells across the TGF-β-treated trajectory had a distinct set of genes from those that control progression through spontaneous EMT, reflecting the direct activation of EMT-associated transcription factors by SMAD signaling[58]. The pseudospatial regions encompassing the first half of the trajectory were strongly enriched for *TGFBR1* and *TGFBR2* knockouts (region 1–4, Fig. 4b

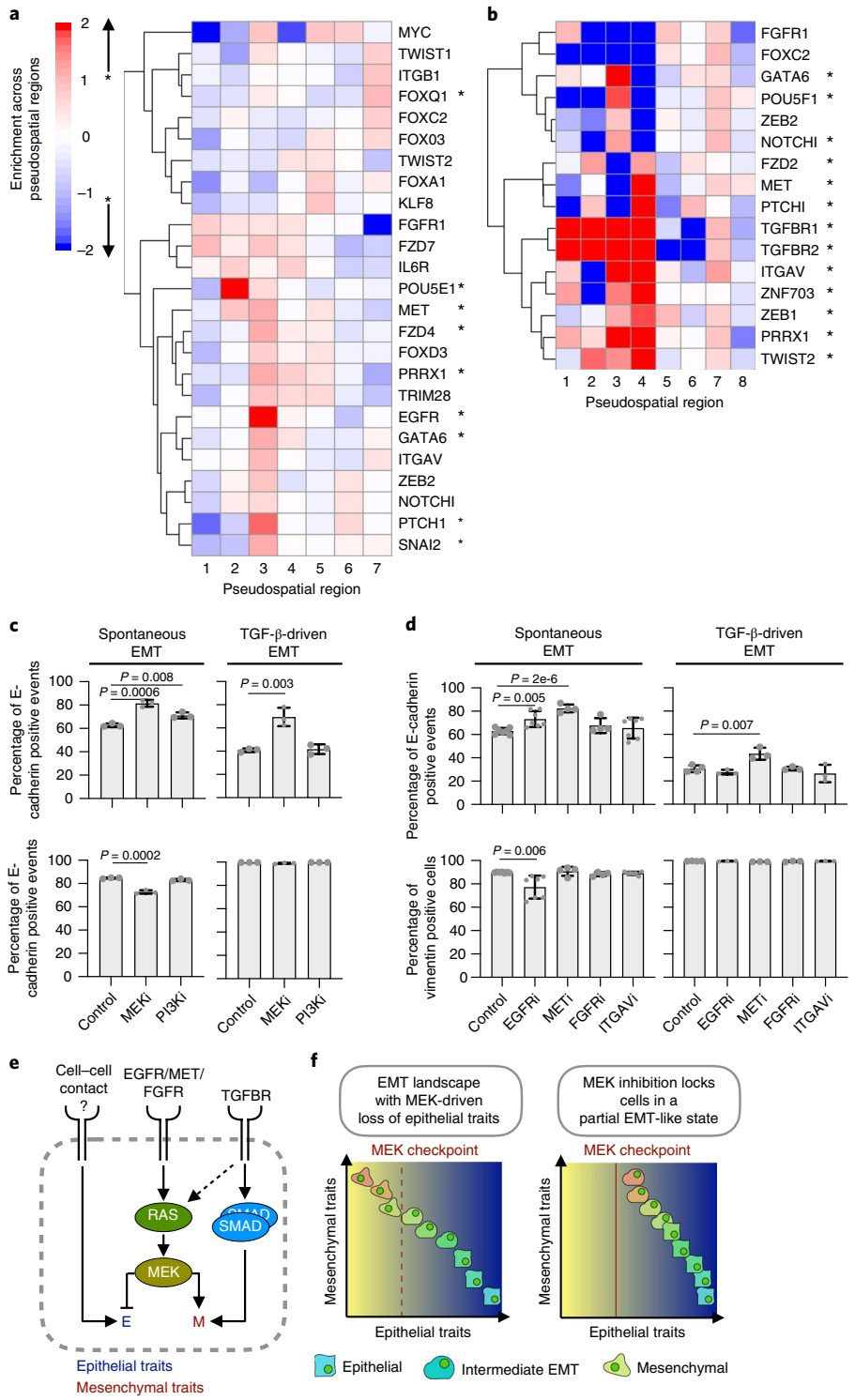


Fig. 4 | Accumulation of knockout cells across spontaneous and TGF-β-driven EMT trajectories identifies regulators of discrete checkpoints across the EMT continuum. a,**b**, Enrichment of knockouts whose distribution is significantly altered across pseudospace, and therefore EMT progression, in our spontaneous (11,908 cells) (**a**) and TGF-β-driven (9,951 cells) (**b**) conditions. The distribution of cells expressing sgRNAs against EMT genes was compared to the distribution of NTC controls by using chi-square (empirically determined FDR < 10%). For targets whose distribution is altered enrichment across each region was determined by calculating the odds ratio. **c**, Percentage of E-cadherin (top panels) or vimentin (bottom panels) positive cells in MCF10A colonies exposed to MEK (U0126) and PI3K (LY294002) inhibition after spontaneous (left panels) or TGF-β-driven (right panels) EMT. Error bars denote standard deviation from the mean ($n = 3$, two-tailed Student's t-test). **d**, Percentage of E-cadherin (top panels) or vimentin (bottom panels) positive cells in MCF10A colonies exposed to EGFR (Erlotinib), MET (Crizotinib), FGFR (Infigratinib) and ITGAV (Cilengitide) inhibition after spontaneous (left panels) or TGF-β-driven (right panels) EMT. Error bars denote standard deviation from the mean (left, spontaneous EMT control/EGFRi/ITGAVi $n = 7$, METi/FGFRi $n = 4$ independent samples; at right: TGF-β-driven EMT control $n = 4$, EGFRi/METi/FGFRi/ITGAVi $n = 3$ independent samples, two-tailed Student's t-test). **e**, Inferred EMT regulatory network and putative regulators identified in this study. **f**, Model depicting the MEK dependent EMT regulatory checkpoint created and its effects on the development of intermediate EMT phenotypes.

and Supplementary Fig. 24b). As in spontaneous EMT, the loss of numerous genes in TGF-β-treated cells concentrated them at defined pseudospatial positions. *ZEB1*, proposed to effect an irreversible switch to a mature mesenchymal state[7], *GATA6*, *NOTCH1* and *POU5F1* were concentrated beginning in region 3, suggesting that this position in the trajectory coincides with a decision point cells pass through during EMT.

Of the seven receptors in our screen known to activate Ras/MAPK signaling, five (*EGFR*, *MET*, *ITGAV*, *ITGB1* and *FGFR1*) altered the distribution of cells over the trajectory, and all but *MET* concentrated them at just one or two regions. Interestingly, only *MET* and *ITGAV* did so during spontaneous and TGF-β-driven EMT. In the spontaneous EMT trajectory, early accumulation of cells expressing sgRNAs against the receptor tyrosine kinases *EGFR* and *MET*[55,56], suggested that one or both are responsible for the early wave of KRAS activity associated with exit from the epithelial state. In the TGF-β-mediated EMT trajectory regions 3 and 4 displayed a robust accumulation of cells expressing sgRNAs against the *ITGAV* integrin and regions 1, 5 and 7 were enriched for cells expressing sgRNAs against the tyrosine kinase *FGFR1*[59]. Integrins function as heterodimers between α and β subunits and αvβ1 heterodimers have been shown to mediate TGF-β signaling during fibrosis[60], a process where EMT has an important role[12,13,61]. These precede the terminal EMT state in our TGF-β trajectory and may contribute to the KRAS-associated late EMT signature identified by our dynamic time warping analysis (Fig. 2c).

To understand how KRAS signaling drives cells through EMT, we performed an in vitro assay in the presence of small molecules that block RAS signaling. RAS exerts its regulatory program via activation of the RAF/MEK/ERK and PI3K/AKT pathways[62,63]. We therefore tested whether loss of MEK (via U0126 treatment) or PI3K signaling (via LY294002 treatment) is sufficient to block the exit from the epithelial state and/or acquisition of mesenchymal phenotypes. Doses of both drugs were chosen to minimize the effects on cell viability (Supplementary Fig. 26). We used flow cytometry to determine the proportion of cells expressing the early EMT markers E-cadherin and vimentin and the mature mesenchymal markers N-cadherin and cytoplasmic fibronectin. On spontaneous EMT, inhibition of PI3K activity led to a modest increase in cells expressing E-cadherin (Fig. 4c). In contrast, MEK inhibition led to a pronounced increase in E-cadherin and an accompanying decrease in vimentin (Fig. 4c,d). Inhibiting MEK prevented downregulation of E-cadherin even in the presence of TGF-β yet had no effect on the proportion of cells expressing vimentin, N-cadherin or fibronectin. Treatment of HuMEC cells with U0126 also decreased the induction of vimentin under spontaneous and TGF-β-driven EMT and decreased fibronectin accumulation after TGF-β exposure (Supplementary Fig. 27).

To map the upstream regulators of this MEK-induced EMT program, we treated MCF10A undergoing spontaneous and TGF-β-driven EMT with small molecules targeting RTKs and integrins from the genetic screen (EGFR-erlotinib, MET-crizotinib, FGFR-infigratinib and ITGAV-cilengitide). Inhibiting EGFR led to an increase in E-cadherin-positive cells and a decrease in vimentin-positive cells only in spontaneous EMT, consistent with EGFR knockout inducing accumulation in pseudospace only in the absence of TGF-β (Fig. 4d). Conversely, MET inhibition led to increases in E-cadherin-positive cells in both spontaneous and TGF-β-driven conditions, reflecting the pausing of knockout cells along both EMT trajectories (Fig. 4d). FGFR and ITGAV inhibition did not significantly alter CDH1 and VIM levels suggesting that they lead to accumulation by alteration of other signaling pathways. We further examined the role of EGFR in regulating the transition into spontaneous EMT by treating cells with a higher dose of erlotinib and expanding our panel of marker proteins. In addition to confirming the regulation of E-cadherin and vimentin by EGFR, we observed that blocking EGFR signaling decreased the level of crumbs3 and desmoplakin during spontaneous EMT (Supplementary Fig. 28a,b). Brightfield images of spontaneous EMT colonies showed a decrease in cells undergoing individual migration, a key phenotypic characteristic of cells transitioning into a mesenchymal state (Supplementary Fig. 28c).

Although inhibiting MEK was not sufficient to prevent activation of the mesenchymal program in MCF10A in the presence of TGF-β, cells coexpressed E-cadherin and high levels of vimentin, N-cadherin and fibronectin protein (Fig. 4c and Supplementary Fig. 29). This suggests that activation of the RAF/MEK/ERK pathway is required for the downregulation of the epithelial program in normal mammary epithelial cells, but that alternate pathways can activate the mesenchymal program when RAF/MEK/ERK signaling is blocked.

Lastly, we explored how the expression of factors that alter the accumulation along EMT in MCF10A relate to the diverse EMT phenotypes observed in HNSCC tumors. Hierarchical clustering of the mean expression level of cell surface receptors identified a strong relationship between receptor expression and the extent of EMT across tumor samples (Supplementary Fig. 30). Expression of *FZD2*, *FZD7*, *FGFR1* and *PTCH1* was inversely correlated with levels of EMT genes. With the exception of *PTCH1*, edited cells lacking these genes were enriched at the beginning of our EMT trajectories (Supplementary Fig. 30). Conversely, tumors expressing high levels of EMT genes (Supplementary Fig. 30) also expressed *MET*, *ITGAV*, *ITGB1*, *TGFBR1* and *TGFBR2*.

Discussion

The integration of single-cell trajectory analysis and pooled genetic screening has the potential to map the genetic circuits that control progression across biological transitions. Understanding the regulation of EMT is a fundamental goal in developmental and cancer biology and has the potential to yield new therapeutic opportunities for intervention in cancer. In contrast to numerous reports of 'partial', 'hybrid' or 'intermediate' EMT stages, both our analysis and recent scRNA-seq and mass cytometry studies of a cancer line[10,11] indicate that cells are organized along a continuum during EMT.

Our CRISPR/scRNA-seq loss-of-function screen reconciles these two conflicting views of gene regulation in EMT. Previously, we showed that a loss-of-function mutation can create a branch from the wild-type trajectory by which cells execute an alternative gene expression program[64]. Here, we show that transcription factor and signaling receptor gene knockouts can cause cells to accumulate at defined points along the EMT continuum, implying the existence of a sequence of 'checkpoints' to progress through it. Therefore, although cells fall along a transcriptional continuum during EMT, genetic insults that disable key signaling pathways could enrich a particular gene expression profile within a cell population, creating the impression of a stable intermediate phenotype. Consistent with this finding, recent single-cell profiling of HNSCC found evidence for diverse partial EMT states at the leading edge of tumors[15], which could arise from genetic heterogeneity amongst cancer cells. Our analysis suggests that local variation in signaling in key pathways could also contribute substantially to the EMT phenotype of a tumor.

Several large modules of genes with distinct wave-like patterns of regulation during spontaneous- or TGF-β-mediated EMT were enriched for targets of KRAS, which may, therefore, be involved throughout the EMT continuum. KRAS signaling can be initiated via various upstream signals, making it difficult to pinpoint the signaling that is driven at each point on the continuum. Focal accumulation of cells lacking particular effectors of KRAS signaling early in spontaneous (*EGFR* and *MET*) and late in TGF-β-mediated (*FGFR2 and ITGAV*) EMT suggests that the cell responds to a sequence of cues to execute steps in the program. TGF-β and RAF/MEK/ERK

are known to be involved in EMT, but how the two pathways interact during the process is not clear. Here, we show that in the absence of exogenous TGF-β, inhibiting RAF/MEK/ERK is sufficient to block exit from the epithelial state and prevent activation of the mesenchymal gene expression program (Fig. 4e). However, when cells are exposed to exogenous TGF-β, this pathway can 'shortcut' MEK to activate the mesenchymal program directly. Further, we find that loss of MEK activity can lock cells in a partial EMT-like state where cells coexpress E-cadherin and high levels of early and late mesenchymal markers. Taken together, these observations point to the existence of 'checkpoints' in the EMT continuum at which cells can arrest and accumulate, creating the impression of discrete stages in bulk cell assays (Fig. 4f).

Our study combines single-cell trajectory analysis with high-throughput pooled loss-of-function screening, which constitutes a powerful approach for identifying upstream signals of pathways that regulate cellular phenotypes. We expect that this methodology will shed light on the genetic architecture that governs not just EMT but diverse biological processes in development and disease. More generally, the observation that interrupting a signaling pathway can enrich a particular transcriptional state within a cell population will inform ongoing debates surrounding the definitions of cell type and state and the delineation of human cellular ontology.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0489-5.

Received: 14 May 2018; Accepted: 23 July 2019;
Published online: 2 September 2019

References

1. Lamouille, S., Xu, J. & Derynck, R. Molecular mechanisms of epithelial-mesenchymal transition. *Nat. Rev. Mol. Cell Biol.* **15**, 178–196 (2014).
2. Nieto, M. A. Epithelial plasticity: a common theme in embryonic and cancer cells. *Science* **342**, 1234850 (2013).
3. Sauka-Spengler, T. & Bronner-Fraser, M. A gene regulatory network orchestrates neural crest formation. *Nat. Rev. Mol. Cell Biol.* **9**, 557–568 (2008).
4. Li, M. et al. Epithelial-mesenchymal transition: an emerging target in tissue fibrosis. *Exp. Biol. Med.* **241**, 1–13 (2016).
5. Nieto, M. A., Angela Nieto, M., Huang, R. Y.-J., Jackson, R. A. & Thiery, J. P. EMT: 2016. *Cell* **166**, 21–45 (2016).
6. Kalluri, R. & Weinberg, R. A. The basics of epithelial-mesenchymal transition. *J. Clin. Invest.* **119**, 1420–1428 (2009).
7. Zhang, J. et al. TGF-β-induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. *Sci. Signal.* **7**, ra91 (2014).
8. Lu, M., Jolly, M. K., Levine, H., Onuchic, J. N. & Ben-Jacob, E. MicroRNA-based regulation of epithelial–hybrid–mesenchymal fate determination. *Proc. Natl Acad. Sci. USA* **110**, 18144–18149 (2013).
9. Hong, T. et al. An Ovol2-Zeb1 mutual inhibitory circuit governs bidirectional and multi-step transition between epithelial and mesenchymal states. *PLoS Comput. Biol.* **11**, e1004569 (2015).
10. Krishnaswamy, S., Zivanovic, N., Sharma, R., Pe'er, D. & Bodenmiller, B. Learning edge rewiring in EMT from single cell data. Preprint at *bioRxiv* https://doi.org/10.1101/155028 (2017).
11. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729.e27 (2018).
12. Grande, M. T. et al. Snail1-induced partial epithelial-to-mesenchymal transition drives renal fibrosis in mice and can be targeted to reverse established disease. *Nat. Med.* **21**, 989–997 (2015).
13. Lovisa, S. et al. Epithelial-to-mesenchymal transition induces cell cycle arrest and parenchymal damage in renal fibrosis. *Nat. Med.* **21**, 998–1009 (2015).
14. Aiello, N. M. et al. EMT Subtype Influences Epithelial Plasticity and Mode of Cell Migration. *Dev. Cell* **45**, 681–695.e4 (2018).
15. Puram, S. V. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611–1624.e24 (2017).
16. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
17. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
18. Scialdone, A. et al. Resolving early mesoderm diversification through single-cell expression profiling. *Nature* **535**, 289–293 (2016).
19. Sarrió, D. et al. Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. *Cancer Res.* **68**, 989–997 (2008).
20. Rodriguez, L. G., Wu, X. & Guan, J.-L. Wound-healing assay. *Methods Mol. Biol.* **294**, 23–29 (2005).
21. Vuoriluoto, K. et al. Vimentin regulates EMT induction by Slug and oncogenic H-Ras and migration by governing Axl expression in breast cancer. *Oncogene* **30**, 1436–1448 (2011).
22. Joost, S. et al. Single-cell transcriptomics reveals that differentiation and spatial signatures shape epidermal and hair follicle heterogeneity. *Cell Syst.* **3**, 221–237.e9 (2016).
23. Schliekelman, M. J. et al. Molecular portraits of epithelial, mesenchymal, and hybrid States in lung adenocarcinoma and their relevance to survival. *Cancer Res.* **75**, 1789–1800 (2015).
24. George, J. T., Jolly, M. K., Xu, J., Somarelli, J. & Levine, H. Survival outcomes in cancer patients predicted by a partial EMT gene expression scoring metric. *Cancer Res.* **77**, 6415–6428 (2017).
25. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* **25**, 25–29 (2000).
26. The Gene Ontology Consortium & The Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic Acids Res.* **47**, D330–D338 (2019)..
27. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
28. Barbie, D. A. et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).
29. Tchernitsa, O. I. et al. Transcriptional basis of KRAS oncogene-mediated cellular transformation in ovarian epithelial cells. *Oncogene* **23**, 4536–4555 (2004).
30. Toivola, D. M., Tao, G.-Z., Habtezion, A., Liao, J. & Omary, M. B. Cellular integrity plus: organelle-related and protein-targeting functions of intermediate filaments. *Trends Cell Biol.* **15**, 608–617 (2005).
31. Huang, R. Y.-J., Guilford, P. & Thiery, J. P. Early events in cell adhesion and polarity during epithelial-mesenchymal transition. *J. Cell Sci.* **125**, 4417–4422 (2012).
32. Feng, Y.-X. et al. Epithelial-to-mesenchymal transition activates PERK-eIF2α and sensitizes cells to endoplasmic reticulum stress. *Cancer Discov.* **4**, 702–715 (2014).
33. Miettinen, P. J., Ebner, R., Lopez, A. R. & Derynck, R. TGF-beta induced transdifferentiation of mammary epithelial cells to mesenchymal cells: involvement of type I receptors. *J. Cell Biol.* **127**, 2021–2036 (1994).
34. CauIln, C., Scholl, F. G., Frontelo, P., Gamallo, C. & Quintanilla, M. Chronic exposure of cultured transformed mouse epidermal cells to transforming growth factor-/1 induces an epithelial-mesenchymal transdifferentiation and a spindle tumoral phenotype. *Cell Growth Differ.* **6**, 1027–1036 (1995).
35. Cacchiarelli, D. et al. Aligning single-cell developmental and reprogramming trajectories identifies molecular determinants of myogenic reprogramming outcome. *Cell Syst.* **7**, 258–268.e3 (2018).
36. Alpert, A., Moore, L. S., Dubovik, T. & Shen-Orr, S. S. Alignment of single-cell trajectories to compare cellular expression dynamics. *Nat. Methods* **15**, 267–270 (2018).
37. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
38. Vintsyuk, T. K. Speech discrimination by dynamic programming. *Cybern. Syst. Anal.* **4**, 52–57 (1968).
39. Rabiner, L. & Juang, B. H. *Fundamentals of Speech Recognition* (PTR Prentice Hall, 1993).
40. Masszi, A. et al. Integrity of cell-cell contacts is a critical regulator of TGF-β1-induced epithelial-to-myofibroblast transition. *Am. J. Pathol.* **165**, 1955–1967 (2004).
41. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
42. Shaul, Y. D. et al. Dihydropyrimidine accumulation is required for the epithelial-mesenchymal transition. *Cell* **158**, 1094–1109 (2014).
43. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
44. Kuo, P.-L., Shen, K.-H., Hung, S.-H. & Hsu, Y.-L. CXCL1/GROα increases cell migration and invasion of prostate cancer by decreasing fibulin-1 expression through NF-κB/HDAC1 epigenetic regulation. *Carcinogenesis* **33**, 2477–2487 (2012).
45. Al-Alwan, L. A. et al. Differential roles of CXCL2 and CXCL3 and their receptors in regulating normal and asthmatic airway smooth muscle cell migration. *J. Immunol.* **191**, 2731–2741 (2013).

46. Tian, X.-J., Zhang, H. & Xing, J. Coupled reversible and irreversible bistable switches underlying TGFβ-induced epithelial to mesenchymal transition. *Biophys. J.* **105**, 1079–1089 (2013).
47. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
48. Dixit, A. et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866.e17 (2016).
49. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882.e21 (2016).
50. Jaitin, D. A. et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-Seq. *Cell* **167**, 1883–1896.e15 (2016).
51. Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Mol. Cell* **66**, 285–299.e5 (2017).
52. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
53. Hill, A. J. et al. On the design of CRISPR-based single-cell molecular screens. *Nat. Methods* **5**, 271–274 (2018).
54. Clark, E. A. & Hynes, R. O. Ras activation is necessary for integrin-mediated activation of extracellular signal-regulated kinase 2 and cytosolic phospholipase A2but not for cytoskeletal organization. *J. Biol. Chem.* **271**, 14814–14818 (1996).
55. Citri, A. & Yarden, Y. EGF–ERBB signalling: towards the systems level. *Nat. Rev. Mol. Cell Biol.* **7**, 505–516 (2006).
56. Peschard, P. & Park, M. From Tpr-Met to Met, tumorigenesis and tubes. *Oncogene* **26**, 1276–1285 (2007).
57. Ornitz, D. M. & Itoh, N. The fibroblast growth factor signaling pathway. *Wiley Interdiscip. Rev. Dev. Biol.* **4**, 215–266 (2015).
58. Xu, J., Lamouille, S. & Derynck, R. TGF-β-induced epithelial to mesenchymal transition. *Cell Res.* **19**, 156–172 (2009).
59. Ahmad, I., Iwata, T. & Leung, H. Y. Mechanisms of FGFR-mediated carcinogenesis. *Biochim. Biophys. Acta* **1823**, 850–860 (2012).
60. Reed, N. I. et al. The v 1 integrin plays a critical in vivo role in tissue fibrosis. *Sci. Transl. Med.* **7**, 288ra79–288ra79 (2015).
61. Kalluri, R. & Neilson, E. G. Epithelial-mesenchymal transition and its implications for fibrosis. *J. Clin. Invest.* **112**, 1776–1784 (2003).
62. Simanshu, D. K., Nissley, D. V. & McCormick, F. RAS proteins and their regulators in human disease. *Cell* **170**, 17–33 (2017).
63. Karnoub, A. E. & Weinberg, R. A. Ras oncogenes: split personalities. *Nat. Rev. Mol. Cell Biol.* **9**, 517–531 (2008).
64. Qiu, X. et al. Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14**, 309–315 (2017).

Acknowledgements

We thank all members of the Trapnell and Shendure laboratories for helpful discussions during the course of this study and feedback on our manuscript, particularly S. Srivatsan, L. Saunders, H. Pliner and J. Packer. We thank N.M. Cruz for feedback on our manuscript. J.L.M.F. thanks S.V. McFaline-Cruz for support. J.L.M.F. was supported by NIH grants T32HL007828 and T32HG000035. A.J.H. was supported by an NSF Graduate Research Fellowship. J.S. and C.T. are supported by NIH grant no. U54DK107979 and the Paul G. Allen Frontiers Group. C.T. is supported by NIH grant nos. DP2HD088158, RC2DK114777 and R01HL118342 and is partly supported by an Alfred P. Sloan Foundation Research Fellowship. J.S. is supported by NIH grant nos. DP1HG007811 and R01HG006283 and is an investigator of the Howard Hughes Medical Institute.

Author contributions

J.L.M.F., J.S. and C.T. devised the project. J.L.M.F., A.J.H., J.S. and C.T. designed experiments. J.L.M.F., A.J.H. and D.J. performed experiments. D.J. and X.Q. provided substantial technical and computational support, respectively. J.L.M.F and A.J.H. performed analyses. J.L.M.F. and C.T. wrote the manuscript with the support of the other authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0489-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Cell culture. MCF10A breast epithelial cells were purchased from ATCC and used within ten passages. HuMEC were purchased from ThermoFisher Scientific and passage 4 cells were used for all experiments. Cas9-expressing MCF10A (MCF10A-Cas9) were generated by transduction with lentiCas9-blast lentivirus (Addgene) and selected with 10 μg ml^{-1} blasticidin (ThermoFisher Scientific) 72 h post-transduction. Cells were cultured at 37 °C and 5% CO_2 in MCF10A media composed of DMEM/F12 (ThermoFisher Scientific) containing 10% fetal bovine serum (ThermoFisher Scientific), 1% Pen-Strep (ThermoFisher Scientific), 10 ng ml^{-1} epidermal growth factor (EGF) (LC Labs), 5 μg ml^{-1} insulin (ThermoFisher Scientific), 10 ng ml^{-1} cholera toxin (List Labs) and 1 μg ml^{-1} hydrocortisone (Sigma).

2D in vitro EMT assay. Before cell seeding the cloning area of a 4.7 mm diameter cloning ring (Sigma) was marked on the bottom of plates and 2.5×10^5 MCF10A, HuMEC or MCF10A-Cas9 cells were seeded within cloning rings placed in the center of the marked well of a six-well tissue culture dish and cells allowed to adhere overnight. Cloning rings were then removed, and wells were washed twice with 3 ml of Dulbecco's PBS (ThermoFisher Scientific) to remove non-adhered cells and MCF10A media added. For TGF-β-treated cells, 4 ng ml^{-1} TGF-β (Pepro Tech) was added to media and TGF-β was replenished every 48 h. Seven days after the cloning rings were removed, inner and outer cell fractions were collected by scraping away outer and inner cells, respectively, using a cell lifter (Costar) and remaining cells were dissociated using TrypLE (ThermoFisher Scientific). The area from which cells were scraped was determined by the outer diameter of the previously marked cloning ring and wells were inspected under a dissecting microscope to assess the purity of the fraction.

Crystal violet and E-cadherin immunofluorescence staining. MCF10A colonies were rinsed with DPBS, fixed by incubating with 4% paraformaldehyde (EM grade, Electron Microscopy Sciences) for 20 min followed by incubation with pure ethanol for 10 min at room temperature. For crystal violet staining, fixed colonies were incubated in 0.05% w/v crystal violet (Sigma) in water for 20 min and excess crystal violet removed by washing 5× for 5 min with DPBS. For E-cadherin staining, fixed colonies were blocked by washing 3× for 5 min in IF buffer (0.1% Triton x-100 (Sigma) and 2% bovine serum albumin (BSA, Fisher Scientific) in PBS). Colonies were then incubated in IF buffer containing mouse anti-E-cadherin antibody (Cell Signaling) for 2 h at room temperature and washed with IF buffer. For imaging, colonies were incubated for 1 h in IF buffer containing Alexa-488 conjugated goat anti-mouse IgG in IF buffer (Invitrogen), washed with IF buffer and 5 μg ml^{-1} Hoechst 33342 (Invitrogen) added to colonies. Brightfield imaging of crystal violet stained whole colonies was performed on a Zeiss Axio Observer by stitching whole well 10× images according to manufacturer's instructions (Carl Zeiss Microimaging). Immunofluorescence imaging of E-cadherin stained colonies was performed by taking representative fields from the center middle and edge of the colony.

Flow cytometry for EMT marker protein levels. MCF10A cells were plated at the center of wells in six-well plates as previously described. Two hours after plating, colonies were washed with PBS and medium was replaced. After 7 d, cells were harvested using TrypLE, washed twice with PBS, resuspended in 500 μl of PBS, fixed by the addition of 5 ml of ice-cold ethanol added dropwise while vortexing and samples stored at −80 °C. Fixed samples were washed and blocked in PBS containing 1% BSA (Sigma). Samples were split into two and one aliquot incubated overnight with mouse anti-cytoplasmic fibronectin antibody (Abcam, ab6328) and rabbit anti-E-cadherin antibody (Cell Signaling, 3195) and the other incubated with rabbit anti-vimentin antibody (Cell Signaling, 5741) and mouse anti-N-cadherin antibody (Cell Signaling, 14215). For spontaneous EMT, cells treated with 500 nM erlotinib (Supplementary Fig. 28), fixed cells were incubated with a mix of rabbit anti-E-cadherin antibody (Cell Signaling, 3195), rat anti-CRB3 antibody (Abcam, ab180835) and mouse anti-desmoplakin I + II antibody (Abcam, ab16434) or a mix of rabbit anti-vimentin antibody (Cell Signaling, 5741) and mouse anti-pan-keratin antibody (Cell Signaling, 4545). Antibody incubations were performed in PBS containing 1% BSA and 0.1% Triton X-100 (Sigma). Samples were washed three times with PBS containing 0.1% Triton X-100, incubated for 1 h with goat anti-rabbit Alexa 647 and goat anti-mouse Alexa-488 secondary antibodies in PBS containing 1% BSA and 0.1% Triton X-100, washed three times with PBS containing 0.1% Triton X-100 and resuspended in PBS for analysis on an LSRII flow cytometer (BD Biosciences) as depicted in Supplementary Fig. 32.

Small-molecule inhibition of KRAS-MEK-ERK pathway activators and flow cytometry for EMT markers. The MEK inhibitor U0126, the PI3K inhibitor LY294002, the EGFR inhibitor erlotinib, the MET inhibitor crizotinib and the FGFR inhibitor infigratinib were purchased from LC Laboratories and resuspended in DMSO. The ITGAV inhibitor cilengitide was purchased from Selleck Chemicals as a 10 mM solution in DMSO. To determine the highest inhibitor concentration that does not have a negative effect on cell viability, we seeded MCF10A cells at 2.5×10^4 cells per well in 96-well plates. After allowing cells to attach overnight wells exposed for 96 h with increasing doses of each inhibitor or DMSO vehicle alone as shown in Supplementary Fig. 23. The highest concentration of each inhibitor that exhibited 90% or higher control of cell growth was used to determine the effect of target inhibition on the induction of a spontaneous and TGF-β-driven EMT. MCF10A and HuMEC cells were plated at the center of wells in six-well plates as previously described. Twenty-four hours after plating, colonies were washed with PBS and cells were pretreated for 1 h in media with or without 1 μM U0126, 1 μM LY294002, 100 nM erlotinib, 1 μM crizotinib, 1 μM infigratininb or 10 μM cilengitide. After preincubation, medium was replaced with medium with or without 4 ng ml^{-1} TGF-β1 as well as any inhibitor with which cells were pretreated. TGF-β1 was replenished every 48 h. After 7 d, samples were harvested and processed for flow cytometry of EMT marker protein levels as described above.

Construction of single-cell RNA libraries and sequencing. Single-cell suspensions of inner and outer cells from Mock and TGF-β-treated MCF10A and HuMEC cells were washed and resuspended in PBS containing 0.04% ultrapure BSA (ThermoFisher Scientific) at 1×10^6 cells per ml. For pseudospatial experiments in the absence or presence of TGF-β presented in Figs. 1 and 2, 2,000–3,000 cells were captured on the Chromium platform (10X Genomics) using one lane per fraction. Single-cell mRNA libraries were built using the single-cell 3' solution V1 kit, libraries sequenced on an Illumina NextSeq 500/550 using 75 cycle high output kits (Read 1 = 64, Read 2 = 5, Index 1 = 14 and Index 2 = 8) and data preprocessed using the Cell Ranger 1.3.1 pipeline (10X Genomics). CROP-seq pseudospatial libraries were generated in a similar fashion capturing 7,000–9,000 cells per fraction. The aggregation option in Cell Ranger was used to normalize libraries to the equivalent number of mean reads per cell specifically: 47,905 and 30,636 mean reads per cell for initial MCF10A and HuMEC pseudospatial experiments, respectively, and 43,557 mean reads per cell for CROP-seq experiments. The percentage of reads mapping to the transcriptome for all samples was between 77.8% and 84.1%. We observed a median of 12,380 and 8,672 unique molecular identifiers (UMI) per cell for initial MCF10A and HuMEC pseudospatial experiments, respectively, and 13,951 median UMIs per cells for CROP-seq experiments. Additional metrics for each individual scRNA-seq library can be found in Supplementary Table 1 and Supplementary Fig. 31.

t-SNE. We performed principal component analysis (PCA) on a matrix composed of cells and gene expression values for genes expressed in more than 50 cells, reduced dimensions to the top 25 principal components and a t-SNE was initialized in this PCA space to reduce to 2 t-SNE dimensions using the reduceDimension function in Monocle2 specifying num_dim = 25, max_component = 2, norm_method = log and reduction_method = t-SNE. To visualize the gene expression level of EMT markers in t-SNE space the gene expression levels of CDH1, DSP and VIM in every cell was normalized by the library size of each cell (the Size_Factor in Monocle2), a pseudocount of 0.1 was added and values $\log_{10}$ normalized.

Pseudospatial reconstruction of single-cell transcriptomes. Trajectories were constructed according to the procedure recommended in the Monocle2 documentation (http://cole-trapnell-lab.github.io/monocle-release/docs/#constructing-single-cell-trajectories). Briefly, genes used to order cells were selected by comparing the inner and outer cell fractions in the assay. For each cell type (MCF10A/HuMEC), differential gene expression analysis was performed between differentialGeneTest() function in Monocle2 (refs. [17,64]). Each gene was fitted to a generalized linear model via the formula 'y ~ cell_fraction', specifying a simple two-group contrast between the fractions. The response (the size-factor adjusted UMIs for the gene) was modeled as a negative binomially distributed random variable. Testing for significant genes was conducted by comparing the model of each gene against a reduced model 'y ~ 1' via a likelihood ratio test.

The top DEGs (likelihood ratio test, FDR, $q < 1 \times 10^{-10}$ and absolute of the $\log_2$ fold-change > 1) were chosen as 'ordering genes' to recover pseudospatial trajectories using the setOrderingFilter(), reduceDimension() and orderCells() functions in Monocle2 using default parameters with the exception of setting ncenter = 500 during dimensionality reduction. Expression of key EMT markers across pseudospace was visualized using the plot_genes_in_pseudotime function in Monocle2 specifying a minimum value of 0.1 (min_expr = 0.1).

Detection and visualization of spatially dependent genes. To extract and visualize genes that vary over a trajectory (beyond the variability one would expect across unordered cells), we used the procedure recommended by the Monocle2 documentation (http://cole-trapnell-lab.github.io/monocle-release/docs/#finding-genes-that-change-as-a-function-of-pseudotime). To identify changes in gene expression across pseudospatial trajectories we fit splines with three degrees of freedom to capture the dynamics of gene expression over psedusopace and tested for differential gene expression analysis using a full model of 'y ~ sm.ns (pseudospace, df = 3)', which encode the position of a cell on the trajectory as a continuous covariate. To further filter genes by those with the largest effect size we divided pseudospace into five quantiles, calculated the AUC for each gene at each quantile and filtered DEGs to those having an AUC > 10 in at least one quantile and an FDR of $< 1 \times 10^{-10}$. DEGs were variance stabilized and scaled, clustered and visualized using the pheatmap function from the R package pheatmap specifying

ward.D2 as the clustering method. To identify biological processes and pathways enriched in clusters of DEGs across pseudospace we performed hypergeometric testing using the piano R package specifying genes expressed in more than 50 cells as the background set.

Calculation of aggregate gene expression scores. To determine the extent to which cells in different samples activate certain gene expression modules we calculated a normalized aggregate expression score for each cell for defined genesets. For a matrix of genes and cells, $\log_{10}$ normalized gene expression was defined for genes in each set after library size normalization and addition of a pseudocount of 1. For each cell, we then calculated the mean normalized expression level of genes in the geneset and mean-centered and variance scaled mean normalized expression values across all cells. The compare_means function from the ggpubr package was used to determine the significance in changes in scores between endpoints in MCF10A and HuMEC EMT trajectories (Supplementary Fig. 13) specifying the wilcox.test as the method and using the Holm procedure (holm) to correct for multiple hypothesis testing.

Dynamic time warping of pseudospatial trajectories. Alignment of Mock and TGF-β-treated trajectories for MCF10A, MCF10A-Cas9 and HuMEC pseudospatial trajectories was performed as described[35], setting the Mock and TGF-β-treated cell trajectories as the reference and query, respectively. Briefly, to arrive at a common pseudospatial axis, trajectories were aligned based on the intersect of genes used for ordering Mock and TGF-β-driven trajectories where pseudospace values were scaled from 0–100, smoothed splines were fitted to each gene using the genSmoothCurves function in Monocle2, splines were variance stabilized and scaled before alignment using the dtw function from the DTW R package using the following options: step pattern = rabiner Juang step pattern (type = 3 and slope.weighting = c), open.begin and open.end = FALSE. To identify genes that describe the differences in the interaction between pseudospace and TGF-β treatment across Mock and TGF-β-driven trajectories we performed differential gene expression analysis using a full model of 'y ~ pseudospace*treatment' and a reduced model of 'y ~ pseudospace'. We isolated DEGs with the largest differences between treatments by dividing pseudospace into five equally spaced quantiles, calculating the AUC (calculated using spline interpolation) for each treatment within each quantile and identifying genes with a relative difference in AUC (relative AUC difference = abs(AUC1 − AUC2)/sum(AUC1 + AUC2)) larger than 0.02 in at least one quantile and an FDR $< 1 \times 10^{-10}$.

Preprocessing of the HNSCC dataset. Processed data from the scRNA-seq of HNSCC tumors described in Puram et al.[15] were downloaded from the GEO Omnibus database (GSE103322) and a Monocle2 Cell Dataset (cds) object was created using gene expression and metadata available in GSE103322_HNSCC_all_data.txt.gz specifying a lower detection limit of 0.1 and choosing tobit as the expression family. Expression values were then converted to mRNA per cell using the Census[64] algorithm implemented in the relative2abs function in Monocle2 after which a new cds object was created specifying negbinomial.size as the expression family. For all analyses, normal cell types and cancer cells from lymph node metastases were excluded. Additionally, only cells that were not processed using the Maxima reverse transcriptase enzyme were chosen for analysis as Puram et al. found that the use of Maxima enzyme could introduce a batch effect. HNSCC tumor samples that had at least 40 cells after applying the exclusion criteria described above were chosen for further analysis.

***k*-nearest-neighbor projection of HNSCC tumor cells onto spontaneous and TGF-β-driven MCF10A EMT trajectories.** We used PCA to reduce the dimensions of a matrix composed of HNSCC tumor cells and MCF10A cells from either spontaneous or TGF-β-driven EMT conditions and gene expression values for genes expressed in more than 50 cells to the top 20 principal components. For each HNSCC tumor sample, we determined the top 20 nearest neighbors from either spontaneous or TGF-β-driven EMT conditions using the *k*-nearest-neighbor search implemented in the FNN R package using the get.knnx function specifying kd_tree as the search algorithm. Finally, the mean pseudospace values for the top 20 MCF10A nearest neighbors were used to assign a pseudospace value for each HNSCC tumor cell.

Reconstruction of HNSCC tumor cell trajectories and alignment to MCF10A EMT trajectories. We chose the four tumor samples with the highest number of cells after applying the exclusion criteria described above (HNSCC17, HNSCC18, HNSCC20 and HNSCC22). Reconstruction of HNSCC trajectories was performed as described above for MCF10A and HuMEC cells with the exception that genes expressed in at least 50 cells across all tumors were used as the feature genes for the setOrderingFilter function in Monocle2. Dynamic time warping of HNSCC and either spontaneous or TGF-β-driven EMT trajectories was performed as previously described with the exception that alignment genes from either spontaneous or TGF-β-driven EMT trajectories were set as alignment genes and the open.begin and open.end parameters of the dtw function in the DTW R package were set to TRUE to allow alignment of HNSCC tumor samples anywhere along the MCF10A trajectories. The dtwPlot function from the DTW R package was used to visualize the alignment of HNSCC trajectories to either MCF10A spontaneous or TGF-β-driven EMT trajectories.

Cloning, lentiviral packaging and transduction of CROP-seq libraries. CROP-seq lentiviral vector (Addgene) was prepped for sgRNA library insertion as described[52]. Briefly, vector was digested using BsmBI (New England Biolabs) and fast alkaline phosphatase (ThermoFisher Scientific). Oligonucleotides (IDT), each containing an sgRNA and homology for Gibson ligation, were designed as follows:

[U6 homology]-[sgRNA]-[sgRNA backbone homology]

5′-tatcttGTGGAAAGGACGAAACACC[G]-[20 bp sgRNA]-gttttagagctaGAA Atagcaagttaaaataagg-3′where the addition of the G immediately upstream of the sgRNA ensures transcription from pol III promoters.

Oligonucleotides (overall design and individual sgRNA sequences can be found in Supplementary Table 10) were made double-stranded by PCR with primers against the invariant regions. The digested CROP-seq vector (10 fmols) and 200 fmols of double-stranded oligonucleotides were ligated using the In-Fusion HD kit (Clontech) by incubation at 50 °C for 1 h. Libraries were then transformed into stellar competent cells (Clontech), transformations were diluted in 250 µl of LB, spread onto 6 LB agar plates containing ampicillin and bacteria culture at 30 °C for 24 h. Resulting colonies were scraped with LB, pooled and vector recovered using a DNA midi kit (Qiagen). Lentivirus was generated by transfecting HEK293T in MCF10A media lacking Pen-Strep with our CROP-seq library using the ViraPower lentiviral packaging mix (ThermoFisher Scientific) according to manufacturer's instructions. Collected lentiviral supernatant was filtered using a 45 µm steriflip vacuum filter (Fisher Scientific). MCF10A-Cas9 cells were transduced with increasing amounts of the CROP-seq lentiviral library and selected with puromycin, retaining transduced cells that had an approximate multiplicity of infection of 0.3.

Enrichment of sgRNA containing transcripts and genotype assignment. For CROP-seq experiments, a nested PCR was performed on 5–10 ng of unsheared cDNA to enrich for sgRNAs positioned on the 3′ UTR of the puromycin resistance gene transcripts. All oligonucleotide sequences used for enrichment of sgRNA containing transcripts can be found in Supplementary Table 10. Briefly, PCR reactions were performed using a P7 reverse primer equivalent to the one introduced by the oligo containing beads in the 10X Chromium Single-cell 3′ solution V1 (5′- CAAGCAGAAGACGGCATACGA -3′). For the first PCR, the forward primer directed toward the beginning of the U6 promoter was:

5′-TTTCCCATGATTCCTTCATATTTGC -3′

For the second PCR, the forward primer binds at the beginning of the sgRNA and adds the standard Nextera R1 sequence:

5′- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGcTTGTGGAAAGG ACGAAACAC -3′

In the final PCR, amplicons were indexed with standard Nextera P5 index primers:

5′-AATGATACGGCGACCACCGAGATCTACAC[10 bp Index] TCGTCGGCAGCGTC -3′

A 1× Ampure cleanup was performed after each PCR. A fifth of PCR1 was added to PCR2 and a 25th of PCR2 was added to PCR3. Libraries were sequenced as spike-ins with transcriptome scRNA-Seq libraries. Final cellular barcodes and UMIs were extracted from position sorted BAM file output by Cell Ranger 1.3.1. We then attempted to find a perfect match for sequences preceding the sgRNA (GTGGAAAGGACGAAACACCG) or used a striped Smith–Watterman alignment to locate the sequence within an error tolerance of 2 bp shorter than the expected sequence. For each match or alignment, the sgRNA sequence is extracted and compared to a whitelist of all sgRNA within an edit distance of half the minimum distance between any pair of guides in our sgRNA library tracking matches for each cell. Chimeric sequences were removed by the approach as detailed in a previous report[48]. sgRNA sequences with over three reads accounting for more than 7.5% of sgRNA reads assigned to a given cell were assigned to each cell. These assignments were combined with the filtered gene expression matrix created by Cell Ranger to assign high-quality cells.

t-SNE and distribution of knockout cells across PCA space. We performed PCA on a matrix composed of cells each containing only one guide from our CROP-seq screen and gene expression values for genes expressed in more than 50 cells and reduced dimensions to 25 principal components. t-SNE was initialized in this PCA space to reduce to two t-SNE dimensions. We then performed louvain clustering across PCA space. A chi-square test was performed to determine whether the distribution of a sgRNA and targets in PCA was significantly different compared to NTC at an FDR cutoff of 5%. Knockouts whose distribution was significantly different from NTC were subjected to further analysis. For each sgRNA we derived a weight to estimate the functional editing rate using an expectation-minimization approach by first modeling the PCA distribution as a mixture of cells with functional and non-functional edits where the mixing parameter is the relative functional edit rate for the sgRNA; estimating the weighted average of the empirical PCA distribution for each guide; and estimating relative functional edit rate as the one that maximizes the observed PCA distribution. Weighted

contingency tables were then generated containing the PCA clusters and weighted cell counts across clusters. Fisher's exact test was used to identify knockouts enriched across PCA clusters. Chi-square and Fisher's exact test were performed using chisq.test and fisher.test functions in R, respectively.

Reproducibility of spatially dependent differential gene expression within our pooled CROP-seq screen. Differential gene expression analysis between isolated cell fractions was performed for cells expressing non-targeting control guide RNAs from our pooled CROP-seq screen under spontaneous and TGF-β-driven EMT conditions using a full model of 'y ~ cell_fraction'. The overlap of DEGs was then compared to initial spatial experiments in MCF10A. The correlation (Pearson's *r*) of beta coefficients between cells from our initial spatial experiments and cells expressing non-targeting control guide RNAs across the full list of original spatial DEGs was calculated using the cor.test function in R.

Trajectory reconstruction of CROP-seq loss-of-function screen and calculating knockout enrichment across pseudospace. Spontaneous and TGF-β-driven EMT trajectories of our CROP-seq loss-of-function screen were individually constructed and aligned as described above. To identify changes in gene expression along pseudospace on gene editing we subsetted cells expressing sgRNAs against a target or non-targeting controls, estimated gene-level dispersions and performed a differential gene expression analysis using a full model of 'y ~ pseudospace*genotype' and a reduced model of 'y ~ pseudospace'. After repeating for all targets, *P* values were corrected for multiple hypothesis testing using the Benjamini–Hochberg procedure. Differential gene expression analysis of the effect of each individual knockout on gene expression over pseudospace is dependent on the rate of non-functional edits, the penetrance of the resulting phenotype and the number of cells expressing sgRNAs against a particular target in our screen. To overcome these challenges and identify changes in the distribution of edited cells along EMT, cells were then clustered according to where cells accumulated along pseudospace into pseudospatial regions along spontaneous and TGF-β pseudospace coordinates using the density peak algorithm[17] implemented in Monocle2 using the clusterCells function. As not all guide RNAs are equally efficient and not all edited cells contain an edit that leads to loss-of-function we used an expectation-maximization approach to estimate the functional edit rate of every guide RNA relative the most efficient guide RNA for a target as described[53]. Briefly, we modeled the distribution of guide RNA containing cells across pseudospatial regions as a mixture of cells with a functional edit and the distribution of cells expressing non-targeting control guide RNAs with the mixing parameter being the functional edit rate for a particular guide RNA. For the expectation step of our model, we estimate the distribution of cells with a functional edit as the weighted average of the relative functional edit rate. Lastly, for the maximization step, we chose the relative function edit rate for a guide RNA as that which maximizes the likelihood of the observed distribution of cells expressing a guide RNA across pseudospatial regions under the mixture model. We then used these guide RNA weights to arrive at weighted cell counts of guide RNA containing cells across pseudospatial regions.

We assessed whether the distribution of guide RNA containing cells as well as a random subset of non-targeting control cells was significantly different compared to the larger pool of non-targeting control expressing cells across pseudospatial regions using a chi-square test. To determine an empirical FDR, we repeated this procedure for 1,000 iterations and calculated the rate at which cells expressing guide RNAs against a target were identified as more significantly distributed across regions compared to the random subset of non-targeting control cells. To obtain a score for the enrichment of differentially distributed targets (FDR < 0.1) across pseudospatial regions, we calculated the odds ratio for each target-region pair using fisher.test in R with the presence or absence of non-targeting control cells in the region and background as failures in our contingency tables. Targets accumulated in at least one region at an enrichment score ($\log_2$ of the odds ratio) of 1 or higher were regarded as strongly enriched. Finally, hierarchical clustering of the enrichment score across pseudospatial regions for differentially distributed targets was used to visualize the accumulation of cells across pseudospace using the pheatmap function in the pheatmap package specifying ward.D2 as the clustering method.

Mean expression of enriched targets in HNSCC tumor cells. Expressed cell surface receptors and transcription factors were identified as those expressed across a minimum of 50 cells in tested HNSCC tumors. To determine the mean expression levels of enriched targets from our CRISPR-Cas9 screen in HNSCC tumors we $\log_{10}$ normalized gene expression for the defined genes after library size normalization and addition of a pseudocount of 1. Then the mean expression level across all cells for a given tumor were averaged. Results were visualized as a heatmap of enriched target gene expression levels using the pheatmap function in the pheatmap R package. Expression profiles were clustered specifying ward.D2 as the clustering method. A column annotation was added depicting the relative partial EMT rank for every tumor as observed by Puram et al.

Statistical methods. Differential gene expression analyses were performed using the differential gene test implemented in Monocle2 and test results corrected for multiple hypothesis testing using the Benjamini–Hochberg procedure. Wilcoxon rank-sum test was used to determine statistical significance of the differences in aggregate gene expression scores for cells across various treatments with correction for multiple hypothesis testing performed using the Holm procedure. Two-tailed Student's *t*-tests were used to determine statistical significance of changes in EMT marker protein expression measured via flow cytometry.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data are available on GEO under accession number GSE114687. Data will also be provided via the Github repository described in 'Code availability'.

Code availability

Code can be found on Github at https://github.com/cole-trapnell-lab/pseudospace.

nature research

Corresponding author(s): Cole Trapnell

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	The single cells analysis pipeline cellranger (version 1.3.0, 10X Genomics) was used to collect the data used in this article.
Data analysis	The single cells analysis package Monocle2 (version 2.6.3) was used in this article. A copy of the analyses performed will be available for distribution on github upon publication.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data is available on GEO under accession number GSE114687 and provided via a Github repository upon publication.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For the initial pseudospatial analysis experiments pertaining to Figures 1 and 2 we determined 2000 cells per fraction to be a reasonable number to robustly identify the underlying trajectory of cells in a well. For single-cell loss of function experiments (Figures 3 and 4), the number of cells for single cell RNA-Seq were determined by obtaining a reasonable amount of coverage in terms of minimum number of cells per target (roughly more than 50 cells per genotype). No statistical methods were used to predetermine sample sizes.
Data exclusions	No data exclusions
Replication	A quantitative comparison of the results of our initial pseudospatial experiments pertaining to Figures 1 and 2 to cells expressing non-targeting control guide RNAs within our pooled screen, pertaining to Figures 3 and 4, identified strong agreement between experiments. Flow cytometry experiments were performed in biological replicate (n = 4-7) and the mean and standard deviation from the mean of the measurements reported. Attempts at replication were successful.
Randomization	Not applicable
Blinding	Not applicable

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Rabbit anti-E-cadherin, Cell Signaling Technologies, #3195 lot 13. Mouse anti-N-cadherin, Cell Signaling Technologies, #14215 lot 2. Rabbit anti-vimentin, Cell Signaling Technologies, #5741 lot 5. Mouse anti-cytoplasmic-fibronectin, Abcam, #ab6328 lot GR3193980-1. Rat anti-CRB3, Abcam, #ab180835 lot GR32532558-1. Mouse anti-desmoplakin I+II, Abcam #ab16434 lot GR3232461-2. Mouse anti-pan-Keratin, Cell Signaling Technologies #4545 lot 1.
Validation	The specificity for all antibodies was confirmed by the manufacturer via immunoblotting confirming that antibodies recognize proteins at the expected molecular weights and via immunofluorosence staining confirming that antibodies recognize proteins with the expected sub-cellular localization. Additionally, anti-E-cadherin and anti-vimentin antibodies were validated for flow cytometry via comparison of e-cadherin low (Hela) and high (MCF7) cell lines and anti-vimentin antibody incubated cells vs. an IgG isotype control, respectively. All antibodies were validated by the manufacturer for specificity to their appropriate antigen.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	MCF10A breast epithelium cells were purchased from ATCC (CRL-10317). Primary human mammary epithelial cells were purchased from Thermo-Fisher Scientific (A10565).

Authentication	MCF10A were not authenticated but used within 10 passages of purchase. All HuMEC experiments were performed with passage 4 cells.
Mycoplasma contamination	MCF10A cell were tested and confirmed negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	Not a commonly misidentified cell line.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	MCF10A and HuMEC cells were seeded in a cloning ring in the center of each well of a 6 well dish, the next day the ring was removed and cells allowed to undergo a spontaneous EMT with or without TGF-B. TGF-B was replenished every 48 hours. After 7 days, cells were harvested by trypsinization, washed twice with PBS, resuspended in 500 uL of cold PBS and 5 mL of ice cold ethanol were added drop-wise to cells while vortexing at low speed. Samples were washed twice with PBS containing 1% BSA (PBS-B) and blocked for 1 hour at room temperature wit PBS-B. Each sample was divided into two, a mix of rabbit anti-e-cadherin/mouse anti-fibronectin or rabbit anti-vimentin/mouse anti-n-cadherin added to one of the two aliquots and samples incubated for 2 hours at room temperature in PBS containing 1% BSA and 0.1% tryton X-100 (PBS-TB). After which, cells were washed twice with PBS-TB, and incubated for 1 hour at room temperature in a mix of Alexa-488 conjugated goat anti-mouse and Alexa-647 conjugated goat anti-rabbit secondary antibodies. Finally, cells were washed twice with PBS-TB and resuspended in PBS for flow cytometric analysis.
Instrument	Data was collected on a BD Bioscience LSRII.
Software	The data was collected using FACSDiva version 8 software. Data was analyzed using FlowJo 10.
Cell population abundance	All expected positive cell populations were present at an abundance of 15% or higher.
Gating strategy	Before analysis of fluorescence, single cells were isolated via sequential gating on SSC-A vs. FSC-A, FSC-H vs FSC-W and SSC-H vs SSC-W according to standard flow cytometry practices. Gates for APC-A (describing e-cadherin or vimentin levels) and FITC-A (describing n-cadherin or fibronectin) were set using the spontaneous EMT sample as a negative control for n-cadherin and fibronectin low populations and the TGFB driven EMT as a negative control for e-cadherin and vimentin low populations.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.