nature methods

ARTICLES

https://doi.org/10.1038/s41592-019-0433-8

MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices

Christopher S. McGinnis[1,10], David M. Patterson[1,10], Juliane Winkler[2], Daniel N. Conrad[1], Marco Y. Hein[3,4], Vasudha Srivastava[1], Jennifer L. Hu[1], Lyndsay M. Murrow[1], Jonathan S. Weissman[3,4], Zena Werb[2,5], Eric D. Chow[6,7*] and Zev J. Gartner[1,5,8,9*]

Sample multiplexing facilitates scRNA-seq by reducing costs and identifying artifacts such as cell doublets. However, universal and scalable sample barcoding strategies have not been described. We therefore developed MULTI-seq: multiplexing using lipid-tagged indices for single-cell and single-nucleus RNA sequencing. MULTI-seq reagents can barcode any cell type or nucleus from any species with an accessible plasma membrane. The method involves minimal sample processing, thereby preserving cell viability and endogenous gene expression patterns. When cells are classified into sample groups using MULTI-seq barcode abundances, data quality is improved through doublet identification and recovery of cells with low RNA content that would otherwise be discarded by standard quality-control workflows. We use MULTI-seq to track the dynamics of T-cell activation, perform a 96-plex perturbation experiment with primary human mammary epithelial cells and multiplex cryopreserved tumors and metastatic sites isolated from a patient-derived xenograft mouse model of triple-negative breast cancer.

Single-cell and single-nucleus RNA sequencing (scRNA-seq and snRNA-seq) have emerged as powerful technologies for interrogating the heterogeneous transcriptional profiles of multicellular systems. Early scRNA-seq workflows were limited to analyzing tens to hundreds of single-cell transcriptomes at a time[1,2]. With the advent of single-cell sequencing technologies based on microwells[3], combinatorial indexing[4,5] and droplet-microfluidics[6–9], the parallel transcriptional analysis of 10^3–10^5 cells or nuclei is now routine. This increase in cell throughput has catalyzed efforts to characterize the composition of whole organs[10] and entire organisms[4,11].

These technologies will increasingly be used to reveal the mechanisms by which cell populations interact to promote development, homeostasis and disease. This shift from descriptive to mechanistic analyses requires integrating spatiotemporal information, diverse perturbations and experimental replicates to draw strong conclusions[12,13]. While existing methods can assay many thousands of cells, sample-specific barcodes (for example, Illumina library indices) are incorporated at the very end of standard library preparation workflows. This practice necessitates the parallel processing of individual samples, which limits scRNA-seq sample-throughput due to reagent costs and the physical constraints of droplet-microfluidics devices. Sample multiplexing approaches address this limitation by labeling cells with sample-specific barcodes before pooling and single-cell isolation. Much as transcripts are linked to cell barcodes during reverse transcription, these techniques assign cells into sample groups by tracking which cells share sample-specific barcodes. Several multiplexing methods have been described that distinguish samples using pre-existing genetic diversity[14], or introduce sample barcodes using either genetic[15–20] or non-genetic[21–23] mechanisms. However, each of these methods has liabilities, including issues with scalability, universality and the potential to introduce secondary perturbations to experiments.

We identified lipid- and cholesterol-modified oligonucleotides (LMOs and CMOs) as reagents that circumvent many of the limitations of other sample multiplexing techniques. We previously described LMO and CMO scaffolds that rapidly and stably incorporate into the plasma membrane of live cells by step-wise assembly[24]. Here, we adapt LMOs and CMOs into MULTI-seq: scRNA-seq and snRNA-seq sample multiplexing using lipid-tagged indices. MULTI-seq localizes sample barcodes to live cells and nuclei regardless of species or genetic background while preserving cell viability and endogenous gene expression patterns.

Results

MULTI-seq overview. MULTI-seq localizes DNA barcodes to plasma membranes by hybridization to an 'anchor' LMO. The 'anchor' LMO associates with membranes through a hydrophobic 5′ lignoceric acid amide. Subsequent hybridization to a 'co-anchor' LMO incorporating a 3′ palmitic acid amide increases the hydrophobicity of the complex and thereby prolongs membrane retention (Fig. 1a). MULTI-seq sample barcodes include a 3′ poly-A capture sequence, an 8-bp sample barcode and a 5′ PCR handle necessary for library preparation and anchor hybridization. Cells or nuclei carry membrane-associated MULTI-seq barcodes into emulsion droplets where the 3′ poly-A domain mimics endogenous transcripts during hybridization to messenger RNA capture beads. Endogenous transcripts and MULTI-seq barcodes are then linked to a common cell- or nucleus-specific barcode during reverse transcription, which

[1]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA, USA. [2]Department of Anatomy, University of California San Francisco, San Francisco, CA, USA. [3]Department of Cellular and Molecular Pharmacology, University of California San Francisco, San Francisco, CA, USA. [4]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [5]Helen Diller Family Comprehensive Cancer Center, San Francisco, CA, USA. [6]Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA, USA. [7]Center for Advanced Technology, University of California San Francisco, San Francisco, CA, USA. [8]Chan Zuckerberg BioHub, University of California San Francisco, San Francisco, CA, USA. [9]Center for Cellular Construction, University of California San Francisco, San Francisco, CA, USA. [10]These authors contributed equally: Christopher S. McGinnis, David M. Patterson. *e-mail: eric.chow@ucsf.edu; zev.gartner@ucsf.edu

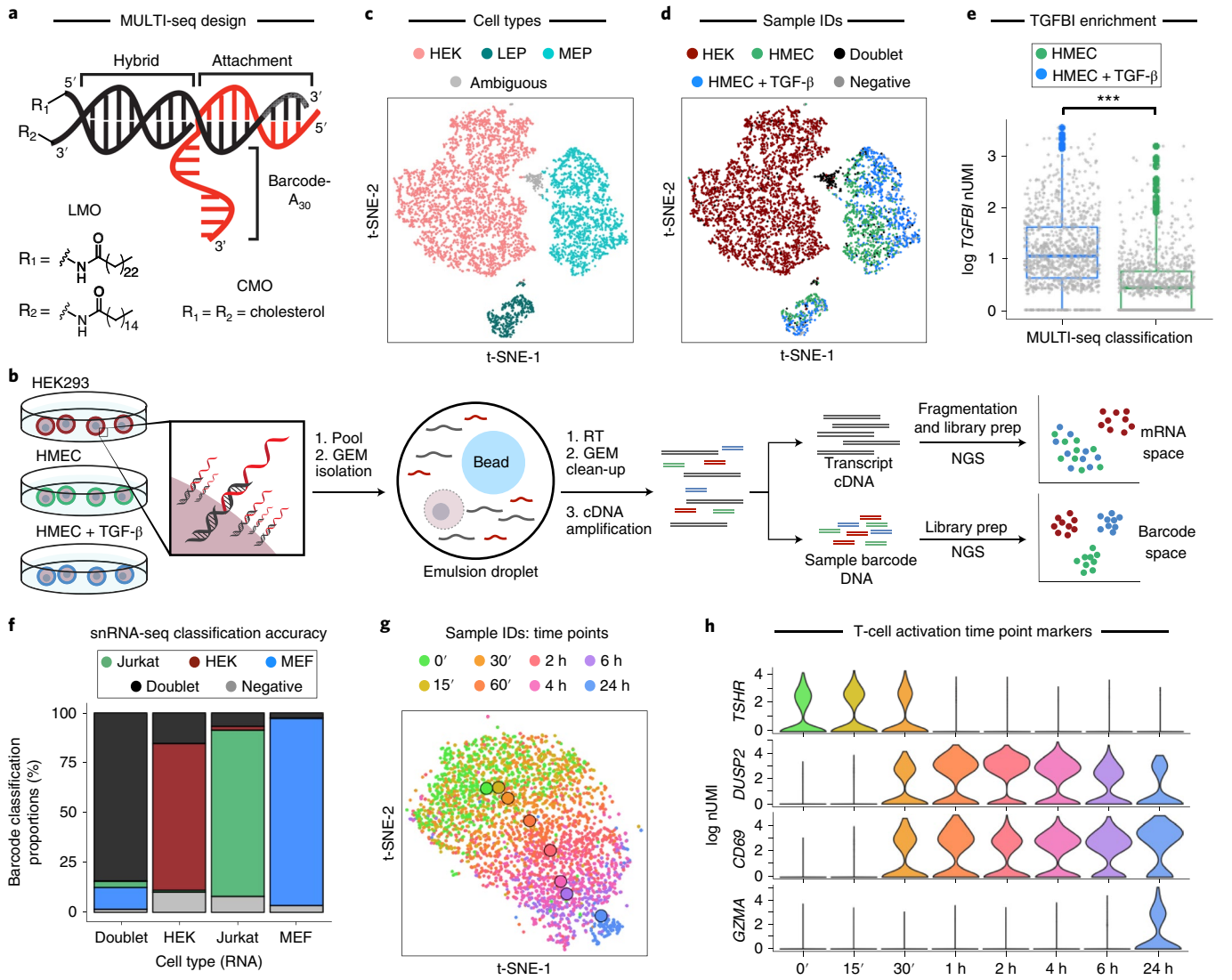

Fig. 1 | MULTI-seq demultiplexes cell types, culture conditions and time-points for single-cell and single-nucleus RNA sequencing. a, Diagram of the anchor/co-anchor LMO and CMO scaffolds (black) with hybridized sample barcode oligonucleotide (red). LMOs and CMOs are distinguished by their unique lipophilic moieties (for example, lignoceric acid, palmitic acid or cholesterol). **b**, Schematic overview of a proof-of-concept single-cell RNA sequencing experiment using MULTI-seq. Three samples (HEKs and HMECs with and without TGF-β stimulation) were barcoded with either LMOs or CMOs and sequenced alongside unlabeled controls. Cells were pooled together before scRNA-seq. NGS produces two UMI count matrices corresponding to gene expression and barcode abundances. GEM, gel bead-in-emulsion; RT, reverse transcription. **c**, Cell-type annotations for LMO-labeled cells: HEKs (pink), MEPs (cyan) and LEPs (dark teal) in gene expression space (see Supplementary Fig. 2a). Ambiguous cells positive for multiple marker genes are displayed in gray. $n = 6{,}186$ MULTI-seq barcoded cells. **d**, MULTI-seq sample classifications for LMO-labeled cells: HEKs (dark red), unstimulated HMECs (green) and TGF-β-stimulated HMECs (blue). Classified doublets (black) predominantly overlap with ambiguously annotated cells. $n = 6{,}186$ MULTI-seq barcoded cells. nUMI, number of unique molecular identifiers. **e**, *TGFBI* expression in TGF-β-stimulated HMECs (blue) and unstimulated HMECs (green). ***Wilcoxon rank-sum test (two-sided), $P <= 10^{-16}$. $n = 1{,}950$ MULTI-seq barcoded HMECs. Data are represented as mean ± s.e.m. **f**, Single-nucleus MULTI-seq sample classification proportions for each cell type identified by clustering in gene expression space (see Supplementary Fig. 2e–g). $n = 5{,}894$ MULTI-seq barcoded nuclei. **g**, MULTI-seq sample classifications in Jurkat cells following activation with ionomycin and PMA for varying amounts of time. Time-point centroids in gene expression space are denoted with larger circles. $n = 3{,}709$ Jurkat nuclei. **h**, Violin plots of gene expression marking different stages of Jurkat cell activation. $n = 3{,}709$ Jurkat nuclei.

enables sample demultiplexing. MULTI-seq barcode and endogenous expression libraries are separated by size selection before next-generation sequencing (NGS) library construction, enabling pooled sequencing at user-defined proportions (Methods). The same strategy can be applied to commercially available CMOs.

We used flow cytometry to evaluate whether LMOs and CMOs predictably label and minimally exchange between live cells at typical sample preparation temperatures of 4 °C (Supplementary Fig. 1a,b). Identical experiments were also performed using freshly isolated nuclei (Supplementary Fig. 1c,d). These data revealed that LMOs exhibit longer membrane residency times than CMOs on live-cell membranes at 4 °C, whereas LMOs and CMOs exchange comparably between live cells at room temperature, suggesting cells should be maintained on ice to achieve optimal sample multiplexing results (Supplementary Fig. 1e). For nuclei, both oligonucleotide conjugates showed minimal exchange between nuclear membranes (Supplementary Fig. 1d); however, bovine serum albumin (BSA) in nuclei isolation buffer specifically quenched LMOs, reducing labeling

efficiency (Supplementary Fig. 1b). While problematic during nuclei labeling, we reasoned that LMO quenching could be strategically employed during live-cell labeling to reduce off-target barcoding and potentially minimize washes before sample pooling. We found that diluting LMO-labeling reactions with 1% BSA in PBS resulted in minimal off-target labeling following pooling (<1% of primary labeling signal), which was 18-fold lower than dilution with PBS (Supplementary Fig. 1f).

MULTI-seq enables scRNA-seq sample demultiplexing. We tested the capacity of MULTI-seq to demultiplex scRNA-seq samples by performing a proof-of-concept experiment using human embryonic kidney (HEK) 293 cells (HEK293) and primary human mammary epithelial cells (HMECs) cultured in the presence or absence of transforming growth factor (TGF)-β (Fig. 1b). Cells were trypsinized, barcoded with LMOs or CMOs and pooled before droplet microfluidic-emulsion with the 10x Genomics Chromium system. In parallel, we prepared unbarcoded replicates to test whether MULTI-seq influenced gene expression or mRNA capture efficiency.

Following data pre-processing (Methods), we analyzed a final scRNA-seq dataset containing 14,377 total cells. We identified clusters in gene expression space according to known markers for HEKs as well as the two cellular components of HMECs, myoepithelial (MEPs) and luminal epithelial cells (LEPs; Fig. 1c and Supplementary Fig. 2a). Projecting MULTI-seq barcode classifications onto gene expression space for LMO-labeled (Fig. 1d) and CMO-labeled cells (Supplementary Fig. 2b) illustrates that both membrane scaffolds successfully demultiplexed each sample. HMECs predicted to have been cultured with TGF-β exhibited enriched TGF-β induced (*TGFBI*) gene expression (Fig. 1e). Moreover, RNA and MULTI-seq barcode unique molecular identifier (UMI) counts were not negatively correlated, demonstrating that MULTI-seq does not impair mRNA capture (Supplementary Fig. 2c). However, we observed transcriptional changes in CMO-labeled HEKs (Supplementary Fig. 2d and Supplementary Table 1) that were absent in LMO-labeled HEKs.

Demultiplexing snRNA-seq and time-course experiments. snRNA-seq is widely used for analyses of solid tissues that are difficult to dissociate[25]. We explored whether MULTI-seq could demultiplex snRNA-seq samples by purifying nuclei from HEKs and mouse embryonic fibroblasts (MEFs) and labeling each pool of nuclei with LMOs or CMOs before snRNA-seq. In parallel, we multiplexed Jurkat cells treated with ionomycin and phorbol 12-myristate 13-acetate (PMA) at eight time-points (0–24 h) to track T-cell activation dynamics (Supplementary Fig. 2e). MULTI-seq sample classifications matched their intended cell-type clusters in gene expression space (Supplementary Fig. 2f,g) with an ~0.5% misclassification rate (Fig. 1f). MULTI-seq classifications were species-specific and predicted ~85% of mouse-human doublets, which approximates the theoretical doublet detection limit for MULTI-seq experiments with 12 samples of 91.7%. Matching live-cell results, MULTI-seq barcoding did not impair mRNA capture (Supplementary Fig. 2h). In contrast to live-cell results, both CMO- and LMO-labeled nuclei were transcriptionally indistinguishable from unbarcoded controls (Supplementary Fig. 2i). Moreover, CMO-labeled nuclei had higher average signal-to-noise ratio (SNR) and total number of barcode UMIs relative to LMO-labeled nuclei (Supplementary Table 2), consistent with previous flow cytometry results.

On demultiplexing individual time-points along the trajectory of T-cell activation (Fig. 1g), we observed multiple literature-supported transcriptional dynamics (Fig. 1h). For example, genes undergoing early down-regulation (for example, *TSHR*[26]) and transient (for example, *DUSP2*, ref. [27]), sustained (for example, *CD69*, ref. [28]) and late (for example, *GRZA*[29]) up-regulation were readily identified in the data.

MULTI-seq identifies doublets in scRNA-seq data. We next sought to demonstrate MULTI-seq scalability by multiplexing 96 unique HMEC samples spanning a range of microenvironmental conditions. We exposed duplicate cultures consisting of MEPs, LEPs and both cell types grown in M87A media[30] without EGF to 15 physiologically relevant signaling molecules[31] or signaling molecule combinations (Supplementary Fig. 3a). We barcoded each sample before pooling and loaded cells across three 10× microfluidics lanes, resulting in a 32-fold reduction in reagent use relative to standard practices.

To classify HMECs into sample groups, we implemented a sample classification workflow inspired by previous strategies[15,16,21] (Methods, Supplementary Materials and Supplementary Fig. 4) that identified 76 sample groups consisting of 26,439 total cells (Supplementary Fig. 3b). Each group was exclusively enriched for a single barcode (Fig. 2a, left, and Supplementary Fig. 3c) an average of ~199-fold above the most abundant off-target barcode (Supplementary Fig. 3d). Unlike sample multiplexing data with relatively few samples, MULTI-seq-defined doublets localized to the peripheries of singlet clusters in barcode space for this experiment (Fig. 2a, right). We suspected that missing barcodes resulted from handling errors during sample preparation (Supplementary Fig. 3b and Supplementary Materials), as a technical replicate yielded all 96 sample groups (Supplementary Fig. 3e–g).

To assess demultiplexing accuracy, we grouped MULTI-seq classifications according to cell-type composition (for example, MEPs alone, LEPs alone or both) and visualized these groups in gene expression space. Unsupervised clustering and marker analysis of the resulting transcriptome data distinguished LEPs from MEPs along with a subset of ambiguous cells expressing markers for both cell types (Fig. 2b, left, and Supplementary Fig. 5a). MULTI-seq classifications matched their expected cell-type clusters (Fig. 2b, right), while cells co-expressing MEP and LEP markers were predominantly defined as doublets. MULTI-seq identified doublets that were overlooked when predicting doublets using marker genes (Fig. 2b, arrow). Additionally, MULTI-seq doublet classifications generally agreed with computational predictions generated using DoubletFinder[32] (Fig. 2c, sensitivity, 0.283 and specificity, 0.965), with the exception of 'homotypic' doublets—that is, doublets formed from transcriptionally similar cells—to which computational doublet detection techniques are insensitive[32,33] (Supplementary Materials). Moreover, DoubletFinder erroneously classified proliferative LEPs as doublets (Fig. 2c, arrow), illustrating how computational doublet inference performance suffers when applied to datasets with low cell-type numbers[32,33].

MULTI-seq identifies transcriptional responses to co-culture conditions and signaling molecules. Sample demultiplexing, doublet removal and quality-control filtering resulted in a final scRNA-seq dataset including 21,753 total cells, revealing two transcriptional responses linked to culture composition. First, we observed that LEPs co-cultured with MEPs exhibited enriched proliferation relative to LEPs cultured alone (Fig. 2d and Supplementary Fig. 5b). In contrast, MEPs were equally proliferative when cultured alone or with LEPs (Supplementary Fig. 5c). Second, we observed that non-proliferative co-cultured MEPs and LEPs were enriched for *TGFBI* expression relative to MEPs and LEPs cultured alone (Fig. 2d, bottom right, and Supplementary Fig. 5d).

We next used hierarchical clustering to assess how LEPs or MEPs responded to signaling molecule exposure. HMECs exposed to the EGFR ligands AREG and EGF exhibited gene expression profiles that were notably different from control cells. AREG- and EGF-stimulated LEPs expressed increased levels of EGFR signaling genes (for example, *DUSP4*, ref. [34]) and genes up-regulated in HER2$^+$ breast cancers (for example, *PHLDA1* (ref. [35]) and Fig. 2e) relative to control LEPs. AREG- and EGF-stimulated MEPs also express high

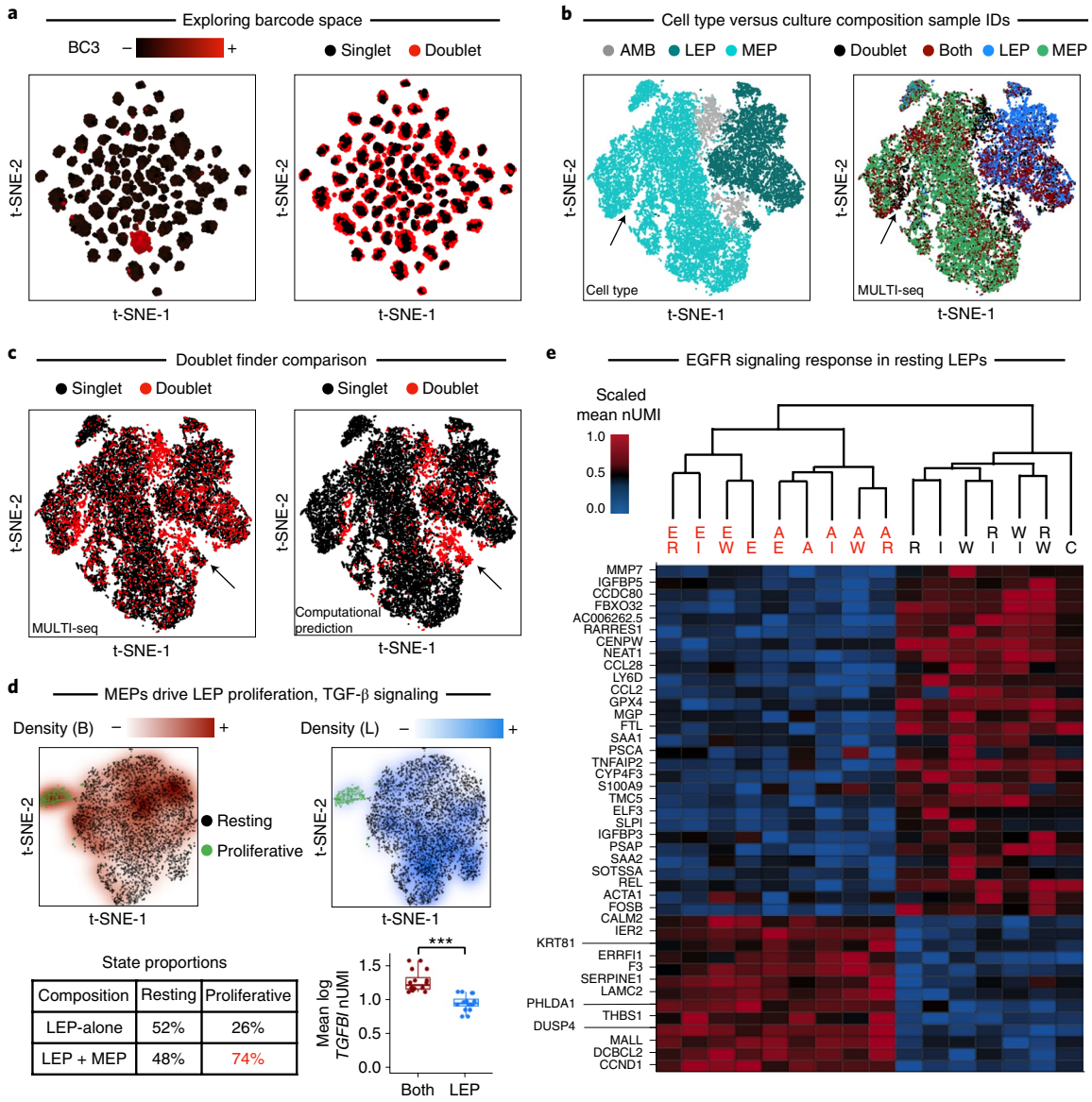

Fig. 2 | MULTI-seq barcoding of multiplexed HMEC culture conditions. a, Barcode UMI abundances (left) and doublet classifications (right) mapped onto barcode space. MULTI-seq barcode number 3 is used as a representative example. Doublets localize to the peripheries of sample groups in large-scale sample multiplexing experiments. $n = 25{,}166$ cells. **b**, Cell-type annotations for MEPs (cyan) and LEPs (dark teal) in gene expression space (left, see Supplementary Fig. 5a). Ambiguous cells positive for multiple marker genes are displayed in gray. MULTI-seq classifications grouped by culture composition (right)—for example, LEP-alone (blue), MEP-alone (green) and both cell types together (dark red)—match cell state annotations. Discordant region where annotated MEPs are classified as MULTI-seq doublets is indicated with arrows. $n = 25{,}166$ cells. **c**, MULTI-seq doublet classifications (left) and computational predictions produced by DoubletFinder (right). Discordant region where DoubletFinder-defined doublets that are classified as MULTI-seq singlets is indicated with arrows. $n = 25{,}166$ cells. **d**, Proliferation and TGF-β signaling in LEPs in MEP-LEP co-culture. Sample classification densities for co-cultured LEPs (B, dark red, top left) and LEPs cultured alone (L, blue, top right) projected onto gene expression space containing resting (black) and proliferative (green) LEPs (see Supplementary Fig. 5b). Resting and proliferating proportions for each culture composition group displayed table (bottom left). Average *TGFBI* expression among LEPs grouped according to growth factor condition (bottom right) reveals co-culture specific enrichment (***Wilcoxon rank-sum test (two-sided), $P = 3.1 \times 10^{-6}$). $n = 32$ signaling molecule condition groups. Data are represented as mean ± s.e.m. **e**, Hierarchical clustering and heat map analysis of resting LEPs grouped by treatment. Emphasized genes are known EGFR signaling targets. RNA UMI abundances are scaled from 0 to 1 for each gene. Values correspond to the average expression within each signaling molecule treatment group. Dendrogram labels: E, EGF; W, WNT4; A, AREG; I, IGF-1; R, RANKL; C, control.

levels of known EGFR-regulated genes (for example, *ANGPTL4* (ref. [36]) and Supplementary Fig. 5e).

MULTI-seq identifies low-RNA cells in cryopreserved, primary patient-derived xenograft (PDX) samples. Using scRNA-seq to analyze archival primary tissue samples is often difficult because these samples can have low cell viability that is compounded during cryopreservation, thawing, enzymatic digestion and scRNA-seq sample preparation. We investigated whether the rapid and non-perturbative nature of MULTI-seq barcoding would enable cryopreserved tissue multiplexing using samples dissected from a PDX mouse model of metastatic triple-negative breast cancer[37]. In this model system, the diameter of primary tumors was used as a proxy for metastatic progression in the lung (Supplementary Fig. 6a).

We barcoded nine distinct samples representing primary tumors and lungs from early- and mid-stage PDX mice (in duplicate), one late-stage PDX mouse and a single lung from an immunodeficient mouse without tumors (Fig. 3a). We then pooled fluorescence-activated cell sorting (FACS)-enriched populations of barcoded hCD298$^+$ human metastases with mCD45$^+$ mouse immune cells before 'super-loading' a single 10x Genomics microfluidics lane.

Quality-control filtering, sample classification and doublet removal resulted in a final scRNA-seq dataset of 9,110 mouse and human singlets spanning all nine samples (Fig. 3b and Supplementary Fig. 6b). Under the conditions tested, barcode SNR was largely invariant to inter-sample differences in total cell number and viability (Supplementary Fig. 6c and Supplementary Table 3). Classification accuracy was supported by tissue-specific gene expression patterns (Supplementary Fig. 6d) and comparisons to FACS enrichment results (Supplementary Fig. 6e). Additionally, MULTI-seq classifications identified high-quality single-cell transcriptomes that would have been discarded using standard quality-control workflows (for example, Cell Ranger RNA UMI inflection point threshold equal to 1,350, Fig. 3c). When comparing cells with 100–1,350 RNA UMIs, classified cells included immune cell types that are difficult to detect using single-cell and bulk transcriptomics (for example, neutrophils[38]). Strikingly, 90.8% of sequenced neutrophils would have been discarded by Cell Ranger. In contrast, unclassified low-RNA cells had poor-quality gene expression profiles[39] (Supplementary Table 4).

Characterizing the lung immune response to metastatic progression. We next sought to describe how lung immune cells respond to metastatic progression. Beginning with a dataset comprising 5,690 mCD45$^+$ cells, we identified gene expression profiles associated with neutrophils, monocytes and macrophages (alveolar, interstitial and (non)-classical monocytes), dendritic cells (mature, immature, Ccr7$^+$ and plasmocytoid DCs) and endothelial cells[10,40] (Fig. 3d, top, and Supplementary Fig. 6f). The use of immunodeficient PDX mice resulted in a lack of lymphocytes (for example, T, B and NK cells).

We observed literature-supported changes in immune cell proportions (Fig. 3d) and transcriptional state (Fig. 3e) at each tumor stage. For instance, neutrophils were enriched in early-stage PDX mice while alveolar macrophages were depleted over the course of metastasis[41,42]. Moreover, stage-specific transcriptional heterogeneity among classical monocytes (Fig. 3f) reflects previous descriptions of lung classical monocyte state transitions in PDX breast cancer models[43].

Unsupervised clustering of classical monocytes cleanly resolved cells from each tumor stage (Supplementary Fig. 6g), enabling the identification of genes up-regulated in classical monocytes during metastatic progression (Supplementary Table 5). Clustering also revealed that classical monocytes from late-stage PDX mice fell into two distinct transcriptional states discernible by *Cd14* expression (Fig. 3f, inset, and Supplementary Table 6) matching previous observations[44]. Genes that are differentially expressed between classical monocyte subsets include genes known to influence metastatic progression[43,45,46] (for example, *Thbs1*, *S100a8*, *S100a9* and *Wfdc21*). To rule out the possibility that these results were primarily due to inter-mouse variability, we used Earth Mover's Distance (EMD)[47] to quantify the magnitude of transcriptional dissimilarity between lung classical monocytes from each mouse and tumor stage. These results illustrate that classical monocytes from early- and mid-stage mouse replicates (scaled EMD = 0.16) were more similar than classical monocytes from distinct tumor stages (scaled EMD = 0.69).

Discussion

MULTI-seq is a scalable multiplexing approach because it uses inexpensive reagents, involves minimal sample handling and is rapid and modular in design. MULTI-seq modularity enables any number of samples to be multiplexed with a single pair of 'anchor' and 'co-anchor' LMOs. Moreover, since LMOs are quenchable with BSA and can be incorporated during proteolytic dissociation, we anticipate that further method optimization will facilitate wash-free sample preparation workflows. When integrated with automated liquid handling, these features position MULTI-seq as a powerful technology enabling 'screen-by-sequencing' applications (for example, L1000 (ref. [48]) and DRUG-seq[49]) in multicellular systems (for example, organoids, PBMCs and so on).

In this study, we leveraged MULTI-seq scalability to perform a 96-plex HMEC perturbation assay, revealing noteworthy principles for future scRNA-seq sample multiplexing experiments. Specifically, we observed that responses to signaling molecules were less pronounced than responses linked to cellular composition. For instance, co-cultured MEPs and LEPs engage in TGF-β signaling that is absent in the associated monocultures. In contrast, MEPs and LEPs only exhibited pronounced transcriptional responses to the EGFR ligands AREG and EGF in these data, despite the established roles of all tested signaling molecules in mammary morphogenesis. We speculate that rich media formulations used to expand cells, such as the M87A media (-EGF) used here, likely buffer cells against microenvironmental perturbations. Thus, careful consideration of cell-type composition and media formulation will be essential to accurately interpret future scRNA-seq experiments.

Beyond its scalability, MULTI-seq improves scRNA-seq data quality in two distinct ways. First, MULTI-seq identifies doublets as cells associated with multiple sample indices. The ability to detect doublets allows for droplet-microfluidics devices to be 'super-loaded', resulting in roughly five-fold improvement in cellular throughput[14,21]. Moreover, unlike computational doublet prediction methods[32,33], MULTI-seq detects homotypic doublets and performs well on scRNA-seq data with minimal cell-type complexity. However, since computational doublet detection methods detect doublets formed from cells with shared sample barcodes[32], doublet detection should ideally involve a synergy of computational and molecular approaches.

Second, MULTI-seq improves scRNA-seq data quality by 'rescuing' cells that would otherwise be discarded by quality-control workflows using RNA UMI thresholds. Such workflows are systematically biased against cell types with low RNA content[39]. MULTI-seq classifications provide an orthogonal metric to RNA UMIs for distinguishing low-RNA from low-quality cells. We leveraged this feature (described initially by Stoeckius et al.[21]) to improve the quality of the PDX dataset, where MULTI-seq classifications 'rescued' >90% of the sequenced neutrophils while avoiding misclassification of broken cells.

Finally, MULTI-seq is universally applicable to any sample including cells or nuclei with an accessible plasma membrane. As a result, we used the same set of MULTI-seq reagents to multiplex 15 distinct cell types or nuclei from both mice and humans. CMOs outperformed LMOs in nuclei isolation buffers containing BSA because BSA sequesters LMOs. Further, we anticipate that MULTI-seq is compatible with sample preservation strategies such as flash-freezing and fixation.

We leveraged all three of these features—scalability, universality and data quality improvement—to multiplex cryopreserved primary tumors and lungs dissected from PDX mouse models at varying stages of metastatic progression. PDX sample multiplexing requires barcoding cells from (1) multiple species that may (2) down-regulate surface epitopes commonly targeted by antibody-based multiplexing techniques (for example, MHC-1, ref. [50]) and (3) have intrinsically low viability requiring minimal sample handling. MULTI-seq successfully demultiplexed every sample, revealing several immune cell responses to metastatic progression in the lung. For example, we confirmed previous reports of metastasis-associated

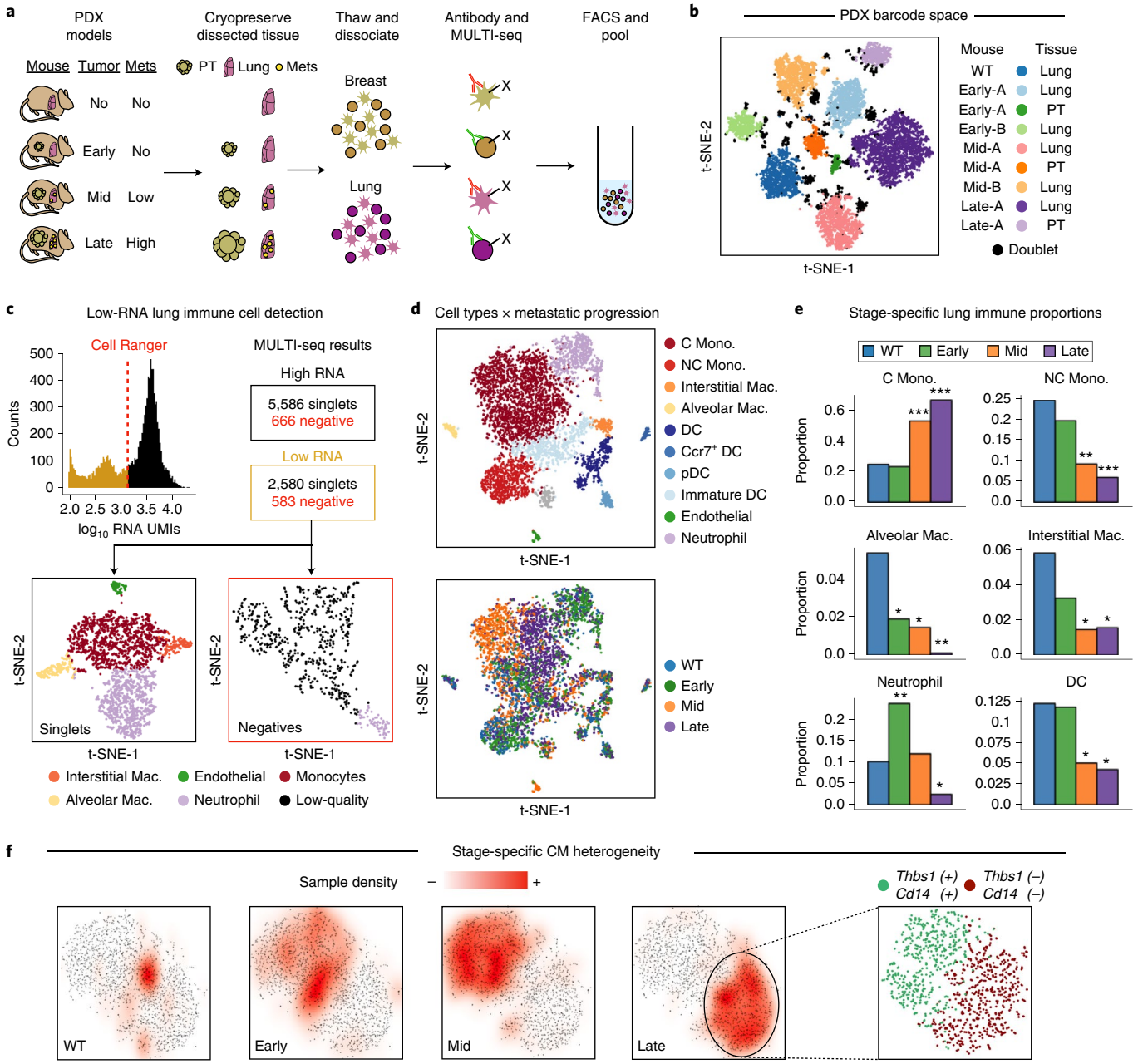

Fig. 3 | PDX sample multiplexing demonstrates low-RNA cell detection, reveals immune cell proportional shifts and classical monocyte heterogeneity in the progressively metastatic lung. a, Schematic overview of PDX experiment. Immune (round) and tumor cells (spindled) isolated from the breast (gold) or lung (pink) were antibody-stained with anti-hCD298 (red) and anti-mCD45 (green) antibodies before MULTI-seq barcoding (X), FACS enrichment, pooling and scRNA-seq. **b**, MULTI-seq sample classifications (wild type (WT), early, mid, late tumor progression) mapped onto barcode space. Replicate tissues are denoted as 'A' or 'B'. $n = 10{,}427$ cells. PT, primary tumor; Mets, metastases. **c**, MULTI-seq classifications facilitate low-RNA and low-quality cell deconvolution. Cell Ranger discards cells barcodes with low RNA UMI counts (red dotted line). Gene expression profiles for classified low-RNA cells reflect established immune cell types (top right; see Supplementary Fig. 6f). Unclassified low-RNA cells resemble low-quality single-cell transcriptomes (bottom right; see Supplementary Table 4). $n = 2{,}580$ (classified), 583 (unclassified) cells. **d**, Cell state annotations (top) and tumor stages (bottom) for lung immune cells in gene expression space. Mono., monocyte; C, classical; NC, non-classical; Mac., macrophage; DC, dendritic cell; pDC, plasmacytoid dendritic cell. Cells with undeterminable annotations displayed in gray. $n = 5{,}965$ cells. **e**, Statistically significant shifts in lung immune cell-type proportions for each tumor stage relative to WT. Two-proportion z-test with Bonferroni multiple comparisons adjustment, $*0.05 > P > 10^{-10}$; $**10^{-10} > P > 10^{-20}$; $***P < 10^{-20}$. $n = 44$ tumor-stage/cell type groups. Statistically insignificant proportional shifts omitted. **f**, Subsetted classical monocyte gene expression space overlaid with sample classification densities corresponding to tumor stage. Inset illustrates heterogeneity within late-stage classical monocytes characterized by differential expression of Thbs1 and Cd14. $n = 2{,}496$ (all), 1,087 (inset) cells.

shifts in neutrophil, alveolar macrophage and classical monocyte proportions. These findings indicate that MULTI-seq can identify known aspects of disease progression. However, we also observed several changes in the immune microenvironment that to our knowledge have not been previously reported, including significant shifts in interstitial macrophages, dendritic cells and non-classical monocytes. Although these findings will require further experimentation for validation, they indicate that MULTI-seq may provide

new insights into disease progression that would be less accessible without sample multiplexing.

Moreover, we identified classical monocyte subsets that were discernible by *Cd14* expression and genes with diverse effects on metastatic progression. Cd14-high classical monocytes expressing the pro-metastatic gene *Thbs1* (ref. [45]) and *Cd14*-low classical monocytes expressing the anti-metastatic genes *S100a8*, *S100a9* and *Wfdc21* (ref. [46]) coexisted in metastasized lungs. Since we isolated immune cells from the whole lung in this study, we could not discern whether *Cd14*-high and *Cd14*-low states were spatially correlated with metastatic sites. However, MULTI-seq could be employed to spatially barcode distinct regions of a single metastatic lung, enabling direct interrogation of classical monocyte spatial heterogeneity.

In summary, MULTI-seq broadly enables users to incorporate additional layers of information into scRNA-seq experiments. We anticipate that in the future, more diverse types of information will be targeted, including spatial coordinates, time-points, species-of-origin and sub-cellular structures (for example, nuclei from multinucleated cells). We also anticipate that increasing LMO membrane residency time using alternative oligonucleotide conjugate designs may enable MULTI-seq applications for non-genetic lineage tracing and/or cellular competition assays.

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/s41592-019-0433-8.

Received: 12 August 2018; Accepted: 29 April 2019;
Published online: 17 June 2019

References

1. Ramsköld, D. et al. Full-length mRNA-Seq from single cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
2. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
3. Gierahn, T. M. et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).
4. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
5. Rosenberg, A. B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
6. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
7. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
8. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
9. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).
10. Tabula Muris Consortium. Single-cell transcriptomic characterization of 20 organs and tissues from individual mice creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
11. Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
12. Wagner, D. E. et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
13. Ordovas-Montanes, J. et al. Allergic inflammatory memory in human respiratory epithelial progenitor cells. *Nature* **560**, 649–654 (2018).
14. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
15. 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).
16. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–82.e21 (2016).
17. Jaitin, D. A. et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* **167**, 1883–1896.e15 (2016).
18. Aarts, M. et al. Coupling shRNA screens with single-cell RNA-seq identifies a dual role for mTOR in reprogramming-induced senescence. *Genes Dev.* **31**, 2085–2098 (2017).
19. Shin, D., Lee, W., Lee, J. H. & Bang, D. Multiplexed single-cell RNA-seq via transient barcoding for simultaneous expression profiling of various drug screening. *Sci. Adv.* **5**, eaav2249 (2019).
20. Guo, C., Biddy, B. A., Kamimoto, K., Kong, W. & Morris, S. A. CellTag indexing: genetic barcode-based sample multiplexing for single-cell technologies. *Genome Biol.* **20**, 90 (2019).
21. Stoeckius, M. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
22. Gehring, J., Park, J. H., Chen, S., Thomson, M. & Pachter, L. Highly multiplexed single-cell RNA-seq for defining cell population and transcriptional spaces. Preprint at https://www.biorxiv.org/content/10.1101/315333v1 (2018).
23. Gaublomme, J. T. et al. Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. Preprint at https://www.biorxiv.org/content/10.1101/476036v1 (2018).
24. Weber, R. J., Liang, S. I., Selden, N. S., Desai, T. A. & Gartner, Z. J. Efficient targeting of fatty-acid modified oligonucleotides to live cell membranes through stepwise assembly. *Biomacromolecules* **15**, 4621–4626 (2014).
25. Wu, H., Kirita, Y., Donnelly, E. L. & Humphreys, B. D. Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis. *J. Am. Soc. Nephrol.* **30**, 23–32 (2019).
26. Coutelier, J. P. et al. Binding and functional effects of thyroid stimulating hormone on human immune cells. *J. Clin. Immunol.* **10**, 204–210 (1990).
27. Jeffrey, K. L. et al. Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1. *Nat. Immunol.* **7**, 274–283 (2006).
28. Ziegler, S. F., Ramsdell, F. & Alderson, M. R. The activation antigen CD69. *Stem Cells* **12**, 456–465 (1994).
29. Lieberman, J. & Fan, Z. Nuclear war: the granzyme A-bomb. *Curr. Opin. Immunol.* **15**, 553–559 (2003).
30. Garbe, J. C. et al. Molecular distinctions between stasis and telomere attrition senescence barriers shown by long-term culture of normal human mammary epithelial cells. *Cancer Res.* **69**, 7557–7568 (2009).
31. Brisken, C. Progesterone signalling in breast cancer: a neglected hormone coming into the limelight. *Nat. Rev. Cancer* **13**, 385–396 (2013).
32. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: doublet detection in single-cell rna sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329–337.e4 (2019).
33. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291.e9 (2019).
34. Chitale, D. et al. An integrated genomic analysis of lung cancer reveals loss of DUSP4 in EGFR-mutant tumors. *Oncogene* **28**, 2773–2783 (2009).
35. Fearon, A. E. et al. PHLDA1 mediates drug resistance in receptor tyrosine kinase-driven cancer. *Cell Rep.* **22**, 2469–2481 (2018).
36. Savage, P. et al. A targetable EGFR-dependent tumor-initiating program in breast cancer. *Cell Rep.* **21**, 1140–1149 (2017).
37. DeRose, Y. S. et al. Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nat. Med.* **17**, 1514–1520 (2011).
38. Jiang, K., Sun, X., Chen, Y., Shen, Y. & Jarvis, J. N. RNA sequencing from human neutrophils reveals distinct transcriptional differences associated with chronic inflammatory states. *BMC Med. Genom.* **8**, 55 (2015).
39. Lun, A. T. L. et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).
40. Reyfman, P. A. et al. Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *Am. J. Respir. Crit. Care Med*. https://doi.org/10.1164/rccm.201712-2410OC (2018).
41. Jablonska, J., Lang, S., Sionov, R. V. & Granot, Z. The regulation of pre-metastatic niche formation by neutrophils. *Oncotarget* **8**, 112132–112144 (2017).
42. Sharma, S. K. et al. Pulmonary alveolar macrophages contribute to the premetastatic niche by suppressing antitumor T cell responses in the lungs. *J. Immunol.* **194**, 5529–5538 (2015).
43. Condamine, T., Ramachandran, I., Youn, J. & Gabrilovich, D. I. Regulation of tumor metastasis by myeloid-derived suppressor cells. *Annu Rev. Med.* **66**, 97–110 (2015).
44. Kitamura, T. et al. Monocytes differentiate to immune suppressive precursors of metastasis-associated macrophages in mouse models of metastatic breast cancer. *Front. Immunol.* **8**, 2004 (2018).
45. Catena, R. et al. Bone marrow-derived Gr1$^+$ cells can generate a metastasis-resistant microenvironment via induced secretion of thrombospondin-1. *Cancer Discov.* **3**, 578–589 (2013).
46. Ouzounova, M. et al. Monocytic and granulocytic myeloid derived suppressor cells differentially regulate spatiotemporal tumour plasticity during metastatic cascade. *Nat. Commun.* **8**, 14979 (2017).
47. Nabavi, S., SChmolze, D., Maitituoheti, M., Malladi, S. & Beck, A. H. EMDomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. *Bioinformatics* **32**, 533–541 (2016).

48. Subramanian, A. et al. A next generation connectivity map: L1000 Platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e17 (2017).
49. Ye, C. et al. DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. *Nat. Commun.* **9**, 4307 (2018).
50. Romero, J. M. et al. Coordinated downregulation of the antigen presentation machinery and HLA class I/beta2-microglobulin complex is responsible for HLA-ABC loss in bladder cancer. *Int. J. Cancer* **113**, 605–610 (2005).

Acknowledgements

This research was supported in part by grants from the Department of Defense Breast Cancer Research Program (nos. W81XWH-10-1-1023 and W81XWH-13-1-0221), NIH (nos. U01CA199315 and DP2 HD080351-01), the NSF (no. MCB-1330864) and the UCSF Center for Cellular Construction (no. DBI-1548297), the 2019 Mary Anne Koda-Kimble Seed Award for Innovation, and the NSF Science and Technology Center. Z.J.G. is a Chan Zuckerberg BioHub Investigator. D.M.P. is supported by the NIGMS of the National Institutes of Health (grant no. F32GM128366). L.M.M. is a Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation (grant no. DRG-2239-15). J.W. and M.Y.H. are supported by EMBO long-term post-doctoral fellowships (grant nos. ALTF-159-2017 and ALTF-1193-2015, respectively). J.L.H. is supported by an NSF GRFP award. We thank M. Thomson for insightful discussions. We thank the UCSF Flow Core (grant no. NIHS10 1S10OD021822-01) and M. Owyong, S. Liu and C. Diadhiou for technical support.

Author contributions

E.D.C. and Z.J.G. conceptualized the method. C.S.M. and D.M.P. designed experiments, synthesized LMOs and optimized the method. C.S.M., D.M.P. and D.N.C. performed analytical flow cytometry experiments. C.S.M. and D.M.P. performed proof-of-concept scRNA-seq experiments. D.M.P. and D.N.C. performed proof-of-concept snRNA-seq experiments. C.S.M. and J.W. performed PDX scRNA-seq experiments. C.S.M., D.M.P., J.L.H. and V.S. performed HMEC scRNA-seq experiments. Z.W. and J.S.W. provided tissue and computational resources, respectively. C.S.M., D.M.P. and L.M.M. performed bioinformatics analysis. C.S.M., M.Y.H., J.W. and J.L.H. implemented the sample classification pipeline. C.S.M. implemented the barcode pre-processing pipeline. C.S.M., D.M.P., Z.J.G. and E.D.C. wrote the manuscript.

Competing interests

Z.J.G., E.D.C., D.M.P. and C.S.M. have filed patent applications related to the MULTI-seq barcoding method. The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0433-8.

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

Correspondence and requests for materials should be addressed to E.D.C. or Z.J.G.

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

Methods

Experimental methods. *Anchor LMO and co-anchor LMO synthesis.* Oligonucleotides were synthesized on an Applied Biosystems Expedite 8909 DNA synthesizer, as previously described (Weber et al.[24], Supplementary Materials).

Cell culture. For the proof-of-concept scRNA-seq and snRNA-seq experiments, HEK293 cells, HMECs, Jurkat cells and MEF cells were maintained at 37 °C with 5% CO_2. HEK293 and MEF cells were cultured in Dulbecco's modified Eagle's medium, high glucose (DMEM H-21) containing 4.5 g l^{-1} glucose, 0.584 g l^{-1} L-glutamine, 3.7 g l^{-1} $NaHCO_3$, supplemented with 10% fetal bovine serum (FBS) and penicillin/streptomycin (100 U ml^{-1} and 100 µg ml^{-1}, respectively). HMECs were cultured in M87A media[30] with or without 24 h of stimulation with 5 ng ml^{-1} human recombinant TGF-β (Peprotech). Jurkat cells were cultured in RPMI-1640 with 25 mM HEPES and 2.0 g l^{-1} $NaHCO_3$ supplemented with 10% FBS and penicillin/streptomycin (100 U ml^{-1} and 100 µg ml^{-1}, respectively).

For the 96-sample HMEC experiments, fourth passage HMECs were lifted using 0.05% trypsin-EDTA for 5 min. The cell suspension was passed through a 45-µm cell strainer to remove any clumps. The cells were washed with M87A media once and resuspended at 10^7 cells ml^{-1}. The cells were incubated with 1:50 APC/Cy-7 anti-human/mouse CD49f (Biolegend, no. 313628) and 1:200 FITC anti-human CD326 (EpCAM) (Biolegend, no. 324204) antibodies for 30 min on ice. The cells were washed once with PBS and resuspended in PBS with 2% BSA with DAPI at 2–4 million cells ml^{-1}. Cells were sorted on BD FACSAria III. DAPI$^+$ cells were discarded. LEPs were gated as EpCAM^{hi}/CD49f^{lo} and MEPs were gated as EpCAM^{lo}/CD49f^{hi} (Supplementary Fig. 7)[51]. This gating strategy results in trace numbers of MEPs and LEPs sorted incorrectly. HMEC sub-populations were sorted into 24-well plates such that wells contained LEPs only, MEPs only or a 2:1 ratio of LEPs to MEPs. Sorted cell populations were cultured for 48 h in M87A media before culturing for 72 h in M87A media (-EGF) supplemented with different signaling molecules or signaling molecule combinations. Specifically, M87A media (-EGF) was supplemented with 100 ng ml^{-1} RANKL, 100 ng ml^{-1} WNT4, 100 ng ml^{-1} IGF-1, 113 ng ml^{-1} AREG and/or 5 ng ml^{-1} EGF (all from Peprotech) alone or in all possible pairwise combinations. For the 96-sample HMEC technical replicate experiment, in vitro cultures were prepared as described above, except all sorted wells contained both LEPs and MEPs. Cultures were then grown in complete M87A media for 72 h before isolation.

scRNA-seq sample preparation. For the proof-of-concept experiment, cells were first treated with trypsin for 5 min at 37 °C in 0.05% trypsin-EDTA before quenching with appropriate cell culture media. Single-cell suspensions were then pelleted for 4 min at 160 relative centrifugal force (rcf) and washed once with PBS before suspension in 90 µl of a 200 nM solution containing equimolar amounts of anchor LMO and sample barcode oligonucleotides in PBS. Anchor LMO-barcode labeling was performed for 5 min on ice before 10 µl of 2 µM co-anchor LMO in PBS (for a final concentration of 200 nM) was added to each cell pool. Following gentle mixing, the labeling reaction was continued on ice for another 5 min before cells were washed twice with PBS, resuspended in PBS with 0.04% BSA, filtered and pooled. The same workflow was also performed with CMOs. LMO-, CMO- and unlabeled control cells were then loaded into three distinct 10x microfluidics lanes.

For the original 96-plex HMEC experiment, LMO labeling was performed during trypsinization to minimize wash steps and thereby limit cell loss and preserve cell viability. HMECs cultured in 24-well plates were labeled for 5 min at 37 °C and 5% CO_2 in 190 µl of a 200 nM solution containing equimolar amounts of anchor LMO and sample barcode oligonucleotides in 0.05% trypsin-EDTA. Then, 10 µl of 4 µM co-anchor LMO in 0.05% trypsin-EDTA was then added to each well (for a final concentration of 200 nM) and labeling/trypsinization was continued for another 5 min at 37 °C and 5% CO_2 before quenching with appropriate cell culture media. A similar labeling protocol was used for the technical replicate experiment, except LMOs were incorporated once the cells were in single-cell suspension. Cells were then transferred to a 96-well plate for washing with 0.04% BSA in PBS. Finally, cells were pooled into a single aliquot, filtered through a 0.45-µm cell strainer and counted before loading 10x microfluidics lanes.

For the PDX experiment, primary tumors and lungs were cryopreserved after dissection from triple-negative breast cancer PDX models generated in NOD-SCID gamma mice as described previously[53]. The UCSF Institutional Animal Care and Use Committee (IACUC) reviewed and approved all animal experiments. On the day of the experiment, cryopreserved tissues were thawed and dissociated in digestion media containing 50 µg ml^{-1} Liberase TL (Sigma-Aldrich) and 2×10^4 U ml^{-1} DNase I (Sigma-Aldrich) in DMEM/F12 (Gibco) using standard GentleMacs protocols. Dissociated cells were then filtered through a 70-µm cell strainer to obtain a single-cell suspension before washing with PBS. Cells were then stained for 15 min on ice with 1:500 Zombie NIR (BioLegend, no. 423105) viability dye in PBS. Cells were then washed with 2% FBS in PBS before blocking for 5 min on ice with 100 µl 1:200 Fc-block (Tonbo, no. 70-0161-U500) in 2% FBS in PBS. After blocking, cells were stained for 45 min on ice with 100 µl of an antibody cocktail containing anti-mouse TER119 (FITC, Thermo Fisher, no. 11-5921-82), anti-mouse CD31 (FITC, Thermo Fisher, no. 11-0311-85), anti-mouse CD45 (BV450, Tonbo, no. 75-0451-U100), anti-mouse MHC-I (APC, eBioscience, no. 17-5999-82) and anti-human CD298 (PE, BioLegend, no. 341704). Cells were then washed with PBS before MULTI-seq labeling for 5 min on ice with 100 µl of 2.5 µM anchor LMO-barcode in PBS. Then, 20 µl of 15 µM co-anchor LMO in PBS was added to each cell pool (for a final concentration of 2.5 µM) and labeling was continued for another 5 min.

We used a ten fold greater LMO concentration for this experiment to account for increases in the total number of cells and lipophilic molecules remaining after dissociation. Following LMO labeling, cells were diluted with 100 µl of 2% FBS in PBS to 'quench' LMOs and washed once in 2% FBS in PBS. Finally, mCD45$^+$ mouse immune cells and hCD298$^+$ human metastases from dissociated primary tumors and lungs were pooled after FACS enrichment, as described previously (Lawson et al.[53], see Supplementary Fig. 8). Cell pools were then processed on a single 10x microfluidics lane.

snRNA-seq sample preparation. For the Jurkat cell activation time-course, 2×10^5 Jurkat cells were added to eight wells of a 12-well plate and treated with 10 ng µl^{-1} phorbol 12-myristate 13-acetate (PMA, Sigma-Aldrich no. P8139) and 1.3 µM ionomycin (Sigma-Aldrich no. I0634) at 15 min, 30 min, 1 h, 2 h, 4 h, 6 h or 24 h before barcoding with LMOs. A single well of Jurkat cells were left untreated. HEK293 and MEF cells were cultured as described above. Nuclei were isolated from cells using a protocol adapted from 10x Genomics. Briefly, suspensions of HEK293, MEF or treated Jurkat cells were washed once with PBS, pelleted at 160 rcf (HEK293, MEF) or 300 rcf (Jurkat) for 4 min at 4 °C and suspended in chilled lysis buffer (0.5% Nonidet P40 Substitute, 10 mM Tris-HCl, 10 mM NaCl and 3 mM $MgCl_2$ in milliQ water) to a density of 2.5×10^6 cells ml^{-1}. Lysis proceeded for 5 min on ice, after which the lysate was pelleted (500 rcf, 4 °C, 4 min) and washed three times in chilled resuspension buffer (2% BSA in PBS). Nuclei were then diluted to a concentration of ~10^6 nuclei ml^{-1} before LMO or CMO labeling. HEK293 and MEF cells were each divided into two samples and labeled with LMOs or CMOs (500 nM in resuspension buffer) using the same procedure as described for live cells (presence of BSA during labeling is the lone alteration as it is required to prevent nuclei clumping). Each Jurkat sample was labeled with LMOs, alone. Each sample was washed three times in 1 ml resuspension buffer (500 rcf, 4 °C, 4 min). The four LMO- and CMO-labeled HEK293 and MEF samples were pooled in equal portions and, separately, Jurkat samples were pooled in equal proportions. These final two samples were combined in a 1:1 ratio and processed on a single 10x microfluidics lane.

scRNA-seq and snRNA-seq library preparation. Sequencing libraries were prepared using a custom protocol based on the 10x Genomics Single Cell V2 and CITE-seq[52] workflows. Briefly, the 10x workflow was followed up until complementary DNA amplification, where 1 µl of 2.5 µM MULTI-seq additive primer was added to the cDNA amplification master mix:

MULTI-seq additive primer: 5′-CCTTGGCACCCGAGAATTCC-3′

This primer increases barcode sequencing yield by enabling the amplification of barcodes that successfully primed reverse transcription on mRNA capture beads but were not extended via template switching (Supplementary Fig. 9c). The MULTI-seq additive primer was erroneously excluded during the proof-of-concept snRNA-seq library preparation and nuclei were still able to be robustly classified. Following amplification, barcode and endogenous cDNA fractions were separated using a 0.6× solid phase reversible immobilization (SPRI) size selection. The endogenous cDNA fraction was then processed according to the 10x workflow until NGS with the formats shown in Supplementary Table 7.

To prepare the barcode fraction for NGS, contaminating oligonucleotides remaining from cDNA amplification were first removed using an established small RNA enrichment protocol (Beckman Coulter). Specifically, we increased the final SPRI ratio in the barcode fraction to 3.2× reaction volumes and added 1.8× reaction volumes of 100% isopropanol (Sigma-Aldrich). Beads were then washed twice with 400 µl of 80% ethanol and allowed to air dry for 2–3 min before elution with 50 µl of Buffer EB (Qiagen). Eluted barcode cDNA was then quantified using QuBit before library preparation PCR (95 °C, 5′; 98 °C, 15′; 60 °C, 30′; 72 °C, 30′; eight cycles; 72 °C, 1′; 4 °C hold). Each reaction volume was a total of 50 µl containing 26.25 µl 2× KAPA HiFi HotStart master mix (Roche), 2.5 µl of 10 µM TruSeq RPIX primer (Illumina), 2.5 µl of 10 µM TruSeq Universal Adapter primer (Illumina), 3.5 ng barcode cDNA and nuclease-free water.

TruSeq RPIX:

5′-CAAGCAGAAGACGGCATACGAGAT**NNNNNN**GTGACTGGAGTT-CCTTGGCACCCGAGAATTCCA-3′

TruSeq P5 adaptor:

5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG CTCTTCCGATCT-3′

Following library preparation PCR, remaining sequencing primers and contaminating oligonucleotides were removed via a 1.6× SPRI clean-up. Representative Bioanalyzer traces at different stages of the MULTI-seq library preparation workflow are documented in Supplementary Fig. 9. Barcode libraries were sequenced using the NGS formats documented in Supplementary Table 2. Sequencing reads predominantly aligned to the barcode reference sequences and resulted in high SNRs with low rates of duplicated UMIs, suggesting that barcode libraries were not sequenced to saturation for any of the presented experiments.

Design and synthesis of LMOs, CMOs and sample barcode oligonucleotides. Anchor and co-anchor LMO and CMO designs were adapted from Weber et al.[24]. Briefly, the anchor LMO has a 5′ lignoceric acid modification with two oligonucleotide domains. The 5′ end is complementary to the co-anchor LMO, which bears a 3′ palmitic acid, and the 3′ end is complementary to the PCR handle of the sample barcode oligonucleotide. The sample barcode was designed to have three components (as in Stoeckius et al.[52]): (1) a 5′ PCR handle for barcode amplification and library preparation, (2) an 8-bp barcode with Hamming distance greater than three relative to all other utilized barcodes and (3) a 30-bp poly-A tail necessary for hybridization to the oligo-dT region of mRNA capture bead oligonucleotides. Identically designed anchor and co-anchor CMOs are conjugated to cholesterol at the 3′ or 5′ ends via a triethylene glycol (TEG) linker and are commercially available from Integrated DNA Technologies.

Anchor: {LA/Chol-TEG}-5′-GTAACGATCCAGCTGTCACTTGGAATTCTCGGGTGCCAAGG-3′

Co-anchor: 5′-AGTGACAGCTGGATCGTTAC-3′-{PA/TEG-Chol}

Sample barcode: 5′-CCTTGGCACCCGAGAATTCCA**NNNNNNNN**A_{30}-3′

Computational methods. *Expression library pre-processing.* Expression library FASTQs were pre-processed using Cell Ranger (10x Genomics) and aligned to the hg19 (proof-of-concept scRNA-seq, HMEC), concatenated mm10-hg19 (PDX) or concatenated mm10-hg19 pre-mRNA (proof-of-concept snRNA-seq) reference transcriptomes. When multiple 10x lanes were sequenced in an experiment, Cell Ranger aggregate was used to perform read-depth normalization.

Cell/nuclei calling. For the proof-of-concept scRNA-seq, snRNA-seq and HMEC technical replicate experiments, cell-associated barcodes were defined using Cell Ranger. For the original 96-plex HMEC experiment, cells were defined as cell barcodes (1) associated with ≥600 total RNA UMIs that (2) were successfully classified during MULTI-seq sample classification workflow. We manually selected 600 RNA UMIs as a threshold to exclude low-quality cell barcodes. For the PDX experiment, we defined cells as barcodes (1) associated with ≥100 total RNA UMIs that (2) were successfully classified during the MULTI-seq sample classification workflow (Supplementary Materials).

Expression library analysis. Following pre-processing and cell/nuclei calling, RNA UMI count matrices were prepared for analysis using the 'Seurat' R package, as described previously[54,55]. Briefly, genes expressed in fewer than three cells were discarded before the percentage of reads mapping to mitochondrial genes (%Mito) was computed for each cell. Outlier cells with elevated %Mito were visually defined and discarded. Data were then $\log_2$ transformed, centered and scaled before variance due to %Mito and the total number of RNA UMIs were regressed out. Highly variable genes were then defined for each dataset by selecting mean expression and dispersion thresholds resulting in ~2,000 total genes. These variable genes were then used during principal component analysis and statistically significant principal components were defined by principal component elbow plot inflection point estimation. Significant principal components were then used for unsupervised Louvian clustering and dimensionality reduction with t-SNE[56].

Following pre-processing, differential gene expression analysis was conducted using the 'FindMarkers' command in 'Seurat', with 'test.use' set to 'bimod'[57] and log fold-change thresholds set in a context-dependent fashion (Supplementary Materials). Other dataset-specific analyses are discussed in the Supplementary Materials. Dataset-specific 'Seurat' pre-processing parameters are given in Supplementary Table 8.

Barcode library pre-processing. Raw barcode library FASTQs were converted to barcode UMI count matrices using custom scripts leveraging the 'ShortRead'[58] and 'stringdist'[59] R packages (Supplementary Fig. 3). Briefly, raw FASTQs were first parsed to discard reads where the first 16 bases of R1 did not perfectly match any of the cell barcodes associated a pre-defined list of cell barcodes. Second, reads where the first eight bases of R2 did not align with <1 mismatch to any reference barcode were discarded. Third, reads were binned by cell barcodes and duplicated UMIs were identified as reads where bases 17–26 of R2 exactly matched. Finally, reference barcode alignment results were then parsed to remove duplicated UMIs before being converted into a final barcode UMI count matrix.

Barcode library sequencing statistics. MULTI-seq barcode library sequencing statistics were computed for classified singlets in all datasets presented in this study. SNR was computed for every cell by finding the quotient of the top two most abundant barcodes. Mean SNRs among all singlets for each dataset presented in this study are documented in Supplementary Table 2. The alignment rate was defined as the proportion of singlet-associated sequencing reads where the first eight bases of R2 aligned with <1 mismatch to any reference barcode.

MULTI-seq sample classification. MULTI-seq barcode UMI count matrices were used to classify cells into sample groups via a workflow inspired by previous scRNA-seq multiplexing approaches[15,16,21] (Supplementary Fig. 3). First, raw barcode reads were $\log_2$-transformed and mean-centered. The presence of each barcode was then visually inspected by performing t-SNE on the normalized barcode count matrix, as implemented in the 'Rtsne' R package with 'initial_dims' set to the total number of barcodes[56]. Missing barcodes (observed only for the 96-plex HMEC experiment) were discerned as those lacking any enrichment in barcode space and were removed.

Next, the top and bottom 0.1% of values for each barcode were excluded and the probability density function (PDF) for each barcode was defined by applying the 'approxfun' R function to Gaussian kernel density estimations produced using the 'bkde' function from the 'KernSmooth' R package[60]. We then sought to classify cells according to the assumption that groups of cells that are positive and negative for each barcode should manifest as local PDF maxima[15,16]. To this end, we computed all local maxima for each PDF and defined negative and positive maxima as the most frequent and highest local maxima, respectively. This strategy assumes that truly barcoded cells will have the highest abundance for any given barcode and that no individual sample group will have more members than the sum of all other groups.

With these positive and negative approximations in hand, we next sought to define barcode-specific UMI thresholds. To find the best inter-maxima quantile for threshold definition (for example, an inter-maxima quantile of 0.5 corresponds to the mid-point), we iterated across 0.02-quantile increments and chose the value that maximized the number of singlet classifications. Sample classifications were then made using these barcode-specific UMI thresholds by discerning which thresholds each cell surpasses, with doublets being defined as cells surpassing >1 threshold[21]. Negative cells (that is, cells surpassing zero thresholds) were then removed and this procedure was repeated until all cells were classified as singlets or doublets. Subsets of negative cells could then be reclassified using semi-supervised learning[21], where singlets defined during the initial workflow are used to initialize cluster centers during *k*-means clustering of negative cells (Supplementary Materials).

Statistical tests. Statistically significant *TGFBI* expression enrichment among TGF-β-stimulated and unstimulated HMECs in the proof-of-concept scRNA-seq experiment was assessed using the Wilcoxon rank-sum test (two-sided, $n = 1{,}950$ cells). Statistically significant *TGFBI* expression enrichment among LEPs and MEPs grouped according to signaling molecule exposure was assessed using the Wilcoxon rank-sum test (two-sided, $n = 32$ signaling molecule condition groups). Differentially expressed genes between clusters in all datasets were defined using the likelihood-ratio test for single-cell gene expression[57] with Bonferroni multiple comparisons adjustment. Statistically significant changes in lung immune cell-type proportions during metastatic progression were assessed using the two-proportion *z*-test with Bonferroni multiple comparisons adjustment ($n = 44$ tumor-stage/cell type groups).

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

Data availability

Raw gene expression and barcode count matrices were uploaded to the Gene Expression Omnibus (GSE129578) along with pertinent metadata.

Code availability

R implementations of the MULTI-seq sample classification and barcode pre-processing pipelines are available in the 'deMULTIplex' R package, and can be downloaded at https://github.com/chris-mcginnis-ucsf/MULTI-seq.

References

51. Lim, E. et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat. Med.* **15**, 907–913 (2009).
52. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
53. Lawson, D. A. et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature* **526**, 131–135 (2015).
54. Satija, R., Ferrel, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
55. 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).
56. van der Maaten, L. J. P. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
57. McDavid, A. et al. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* **29**, 461–467 (2013).
58. Morgan, M. et al. ShortRead: a Bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**, 2607–2608 (2009).
59. van der Loo, M. The stringdist package for approximate string matching. *R J.* **6**, 111–122 (2014).
60. Wand, M. P. & Jones, M. C. *Kernel Smoothing* (Chapman & Hall, 1995).

nature research

Corresponding author(s): Zev Gartner
Eric Chow

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	No software was used data data collection
Data analysis	scRNA-seq data was analyzed primarily using the 'Seurat' R package (V2.3, https://satijalab.org/seurat/). Additional R packages used for scRNA-seq data analysis: DoubletFinder (V2.0, https://github.com/chris-mcginnis-ucsf/DoubletFinder) and EMDomics (V3.8, https://www.bioconductor.org/packages/release/bioc/html/EMDomics.html). scRNA-seq expression library FASTQs were pre-processed using CellRanger (V2.2, 10X Genomics). MULTI-seq sample classification and FASTQ pre-processing was performed the 'deMULTIplex' R package (V1.0, https://github.com/chris-mcginnis-ucsf/MULTI-seq). All figures were made using the 'ggplot2' (V3.1) and 'Seurat' (V2.3) R packages. Analytical flow cytometry analysis was performed using FlowJo.

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

Raw gene expression and MULTI-seq sample barcode count matrices were uploaded to the Gene Expression Omnibus (GSE…). Relevant count and metadata for each main text and supplemental figure are available in the Supplementary Materials.

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	Sample sizes were not selected a priori. Instead, single-cell transcriptomes passing quality-control filtering were utilized to demonstrate key aspects of MULTI-seq methodology performance. Biological interpretations of single-cell RNA sequencing data was constrained by statistical significance.
Data exclusions	No data excluded
Replication	Experiments presented in the paper were not repeated.
Randomization	Randomization is not relevant to this study because we sought to specifically test whether single-cell transcriptome data could be linked to pre-defined MULTI-seq barcodes in a fashion matching expectations. Moreover, sample barcoding was utilized to explore unknown facets of how the cellular responses to perturbations manifest transcriptomically. Thus, randomizing sample barcodes across perturbations would defeat the purpose of our experiments.
Blinding	Blinding was not relevant for this study because technology development requires ground-truth benchmarks with which to assess assay accuracy.

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

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	Triple negative breast cancer PDX models were established and provided by the lab of A. Welm (Derose et al., 2011). These labs provide material transfer upon request.

Antibodies

Antibodies used

Analytical:
None

HMEC: (Antibody Name: Supplier, Catalog Number, Clone)
1:50 APC/Cy-7 anti-human/mouse CD49f: Biolegend, #313628, GoH3
1:200 FITC anti-human CD326 (EPCAM): Biolegend, #324204, 9C4

PDX: (Antibody Name: Supplier, Catalog Number, Clone)
1:200 Fc-block: Tonbo, #70-0161-U500, 2.4G2
1:100 FITC anti-mouse TER119: ThermoFisher, #11-5921-82, TER-119
1:25 FITC anti-mouse CD31: ThermoFisher, #11-0311-85, 390
1:20 (tumor), 1:80 (lung) BV450 anti-mouse CD45: Tonbo, #75-0451-U100, 30-F11
1:40 (tumor), 1:160 (lung) APC anti-mouse MHC-I: eBioscience, #17-5999-82, 28-14-8
1:80 PE anti-human CD298: BioLegend, #341704, LNH-94

Validation

Analytical:
None

HMEC:
Both antibodies were validated and quality control tested for flow cytometry applications.
Antibodies were used to sort MEPs and LEPs from bulk HMECs, as described previously (Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nat Med. 2009; 15(8):907-13.)

PDX:
All antibodies were validated and quality control tested for flow cytometry applications.
Antibodies were used to enrich for human tumor cells and tumor-associated mouse immune cells, as described previously (1. Lawson DA, Bhakta NR, Kessenbrock K, Prummel KD, Yu Y, Takai K, et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. Nature. 2015; 526(7571):131-5.)

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

HEK293: ATCC CRL-1573
MEF: ATCC
Jurkat: ATCC TIB-152
HMEC: Lawrence Berkeley National Laboratory Human Mammary Epithelial Cell (HMEC) Bank

Authentication

HMEC: Primary cells do not require authentication
MEF/Jurkat/HEK293: Cells were authenticated at their source (e.g., ATCC) prior to acquisition, but no extra authentication was employed in this study. Single-cell gene expression profiles match expectations from literature-supported marker genes for each cell line.

Mycoplasma contamination

Cell lines were tested for mycoplasma contamination at their source (e.g., ATCC), but no extra testing was employed in this study.

Commonly misidentified lines
(See ICLAC register)

HEK cells were used in the proof-of-concept single-cell RNA sequencing experiment and analytical flow cytometry experiments because biological interpretation of these datasets was secondary to analysis of lipid-modified nucleotide behavior in a cellular context. Jurkat cells were used in the proof-of-concept single-nucleus RNA sequencing experiment due to their well-characterized temporal response to PMA and Ionomycin stimulation.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

Species: Mus musculus
Strain: NOD-scid gamma
Sex: Female
Age: 5-8 months

Wild animals

The study did not involve wild animals.

Field-collected samples

The study did not involve field-collected samples

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	Analytical: HEKs were trypsinized for 5 minutes at 37°C in 0.05% trypsin-EDTA before quenching with appropriate cell culture media. Single-cell suspensions were then pelleted for 4 minutes at 160 xg and washed once with PBS before resuspension in 90uL of a 200nM solution containing equimolar amounts of anchor LMO and sample barcode oligonucleotides in PBS. Anchor LMO-barcode labeling was performed for 5 minutes on ice before 10 uL of 2uM co-anchor LMO in PBS was added to each cell pool. Following gentle mixing, the labeling reaction was continued on ice for another 5 minutes before cells were washed twice with PBS, resuspended in PBS with 0.04% BSA. HMEC: Fourth passage HMECs were lifted using 0.05% trypsin+EDTA for 5 minutes. The cell suspension was passed through a 0.45 um cell strainer to remove any clumps. The cells were washed with M87A media once and resuspended at 10^7 cells/mL. The cells were incubated with 1:50 APC/Cy-7 anti-human/mouse CD49f (Biolegend, #313628) and 1:200 FITC anti-human CD326 (EpCAM) (Biolegend, #324204) antibodies for 30 minutes on ice. The cells were washed once with PBS and resuspended in PBS with 2% BSA with DAPI at 2-4 million cells/mL prior to FACS PDX: Cryopreserved tissue was dissociated in digestion media containing 50 ug/mL Liberase TL (Sigma-Aldrich) and 2x10^4 U/mL DNase I (Sigma-Aldrich) in DMEM/F12 (Gibco) using standard GentleMacs protocols. All samples were filtered through a 70 um filter to ensure single cell suspensions and were stained on ice for FACS sorting with Zombie NIR (BioLegend, #423105) 15 min in PBS, following antibody staining for 45 min in PBS/2%FBS with Fc-block (Tonbo, #70-0161-U500), anti-mouse TER119 (FITC, ThermoFisher, #11-5921-82), anti-mouse CD31 (FITC, ThermoFisher, #11-0311-85), anti-mouse CD45 (BV450, Tonbo, #75-0451-U100), anti-mouse MHC-I (APC, eBioscience, #17-5999-82) and anti-human CD298 (PE, BioLegend, #341704). MULTI-seq labeling was performed using 100uL of a 2.5uM solution containing equimolar amounts of anchor LMO and sample barcode oligonucleotides in PBS. LMO labeling was performed for 5 minutes on ice before 20uL of 15uM co-anchor LMO in PBS was added to each cell pool. LMO labeling was continued for another 5 minutes on ice before cells were washed once with PBS containing 2% FBS.
Instrument	Analytical: BD FACScalibur HMEC: BD FACS Aria III PDX: BD FACS Aria II
Software	FACS Data Collection Software Analytical: CellQuest HMEC and PDX: BD FACS Diva FACS Data Analysis Software Analytical, HMEC, and PDX: FowJo and R
Cell population abundance	Analytical: n/a HMEC: Fourth-passage HMECs primarily contain two epithelial cell types – luminal and myoepithelial cells (LEP and MEP respectively). The LEP and MEP populations were isolated by gating on cell surface markers – EpCAM and CD49f. LEP and MEP were defined as EpCAMhi, CD49flo and EpCAMlo, CD49fhi respectively. Typically, LEP comprise of 10-15% of the unsorted cells. Post-sort purity was assessed by immunostaining for lineage-specific markers keratin 19 for LEP and p63 for MEP. LEP sorts were ~90% pure whereas MEP sorts were ~98% pure.
Gating strategy	Analytical: Live cells were distinguished from debris and cell aggregates via FSC-A x SSC-A gating. HMEC: Gating on FSC-A x SSC-A and FSC-A x FSC-W was used to eliminate cell aggregates and ensure the collection of only single cells. Live cells were gated as DAPI-. LEP and MEP were defined as EpCAMhi, CD49flo and EpCAMlo, CD49fhi respectively. PDX: FSC-H x FSC-W and SSC-A x SSC-W gating was used to eliminate cell aggregates and ensure single cell sorting. After live-cell enrichment through eliminating NIR+ cells, we sorted live mouse CD45+ and human tumor cells (hCD298+ Lin- mMHC-I-) which excluded contaminating human or mouse haematopoietic and endothelial cells by gating out Lin+ (Ter119, CD31) cells. Positive and negative populations were defined through unstained and FMO controls.

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