Pan-cancer Drivers are Recurrent Transcriptional Regulatory Heterogeneities in Early-stage Luminal Breast Cancer

Shambhavi Singh[1], Matthew D. Sutcliffe[1], Kathy Repich[2], Kristen A. Atkins[3], Jennifer A. Harvey[2,4], Kevin A. Janes[1,5]

[1]Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia.
[2]Department of Radiology, University of Virginia, Charlottesville, Virginia.
[3]Department of Pathology, University of Virginia, Charlottesville, Virginia.
[4]Department of Imaging Sciences, University of Rochester Medical Center, Rochester, New York.
[5]Biochemistry & Molecular Genetics, University of Virginia, Charlottesville, Virginia.

Corresponding Author: Kevin A. Janes, University of Virginia, Box 800759 Health System, Charlottesville, VA 22908. Phone: 434-243-2126; Fax: 434-982-3870; E-mail: kjanes@virginia.edu

The authors declare no potential conflicts of interest.

Running title: Stochastic profiling of luminal breast cancer by 10cRNA-seq

Abstract

The heterogeneous composition of solid tumors is known to impact disease progression and response to therapy. Malignant cells coexist in different regulatory states that can be accessed transcriptomically by single-cell RNA sequencing, but these methods have many caveats related to sensitivity, noise, and sample handling. We revised a statistical fluctuation analysis called stochastic profiling to combine with 10-cell RNA sequencing, which was designed for laser-capture microdissection (LCM) and extended here for immuno-LCM. When applied to a cohort of late-onset, early-stage luminal breast cancers, the integrated approach identified thousands of candidate regulatory heterogeneities. Intersecting the candidates from different tumors yielded a relatively stable set of 710 recurrent heterogeneously expressed genes (RHEGs) that were significantly variable in >50% of patients. RHEGs were not confounded by dissociation artifacts, cell cycle oscillations, or driving mutations for breast cancer. Rather, we detected RHEG enrichments for epithelial-to-mesenchymal transition genes and, unexpectedly, the latest pan-cancer assembly of driver genes across cancer types other than breast. Heterogeneous transcriptional regulation conceivably provides a faster, reversible mechanism for malignant cells to sample the effects of potential oncogenes or tumor suppressors on cancer hallmarks.

Statement of significance

Profiling intratumor heterogeneity of luminal breast carcinoma cells identifies a recurrent set of genes suggesting sporadic activation of pathways known to drive other types of cancer.

Introduction

Approximately 80% of human tumors are epithelial carcinomas (1). Epithelial cells and progenitors proliferate considerably during normal tissue development, maintenance, and repair (2,3). Proper epithelial organization is enforced by basement membrane extracellular matrix (ECM), which becomes compromised during the epithelial cell-state changes underlying carcinomagenesis (4-6). Advanced carcinomas show considerable cell-to-cell variation in chromosomal gains–losses (7), overall mutational burden (8), and hybrid epithelial-mesenchymal traits (9). However, the state trajectory of carcinoma cells within large, rapidly progressing tumors is not stereotyped (10), complicating general interpretations of this variability. It is not known when intra-carcinoma cell heterogeneity meaningfully emerges, nor whether there might be common themes early in tumorigenesis that go on to diverge at later stages.

Thoroughly deconstructing intratumor heterogeneity requires transcriptomic approaches that can separate lineages and distinguish regulatory states with high cellular resolution (11,12). Conventional methods for single-cell RNA sequencing (scRNA-seq) either a) profile dozens–hundreds of cells at the maximum depth afforded by the single-cell sample, or b) shallowly scan (tens of) thousands of cells for molecular phenotyping (13). Regardless of the output, all leading approaches dissociate tumors into single-cell suspensions, requiring up to an hour of tissue processing and yielding a range of carcinoma proportions depending on cancer type (**Table 1**). The impact of these preparative steps on the transcriptomes of live cells is recognized (14,15), but they are considered to be an unavoidable tradeoff of the scRNA-seq approach.

Considering the drawbacks, we (16) and others (17,18) have developed approaches orthogonal to scRNA-seq that combine high-sensitivity transcriptomics with laser-capture microdissection (LCM). Using LCM, histologically distinguishable cell types can be isolated with single-cell resolution from cryosections of snap-frozen tissue or tumor. Unfortunately, all LCM-based sequencing methods have substantially decreased sensitivity and technical reproducibility with fewer than 10 cells of microdissected material (17-19). To gain single-cell information, we reasoned that our 10-cell RNA sequencing (10cRNA-seq) method could productively identify carcinoma cell-regulatory heterogeneities if it were implemented as a type of fluctuation analysis called stochastic profiling (20). Previously, stochastic profiling was applied to 10-cell averaged gene-expression profiles collected repeatedly within a clonal cell line and measured by qPCR or microarray, leading to predicted single-cell heterogeneities that were

subsequently validated (20-22). The distinct statistical properties of RNA sequencing data required design of a customized analytical pipeline for stochastic profiling by 10cRNA-seq.

Here, we combine 10cRNA-seq with an abundance-dependent overdispersion statistic that enables stochastic profiling of tumor cells *in situ.* Without any sample dissociation, we deeply profile carcinoma cell-to-cell heterogeneity in a cohort of five closely matched, late-onset and early-stage luminal breast cancers (**Table 1** and Supplementary Table ST1), obtaining data on 21,255 genes from 1400 carcinoma cells. The LCM component of 10cRNA-seq proved critical to obtain carcinoma heterogeneity profiles from cases with extensive immune infiltration. 10-cell pooling minimized the contribution of periodic transcripts that covary with cell-cycle phases. Stochastic profiling inferred 710 candidate transcripts that were recurrently heterogeneous in ≥50% of tumors. Subsampling tumors in the cohort consistently yielded 500–1500 candidates by the same criteria, suggesting bounds for reliable variability between similar tumors. The shared set of candidates was largely devoid of detachment-induced artifacts (14) and, surprisingly, breast-cancer driver genes (23). Recurrent heterogeneities were instead enriched in the collagen and matricellular constituents of a pan-cancer epithelial-to-mesenchymal transition (EMT) signature (24). The heterogeneities uncoupled canonical EMT marker genes that are tightly correlated at the population level. Most intriguing was an enrichment for dozens of non-cycling driver genes for cancers other than breast (25). Our results raise the possibility that early-stage luminal breast cancers sample a much broader landscape of oncogenes and tumor suppressors through transcriptional heterogeneity than indicated by the genomic lesions characteristic of the subtype.

Materials and Methods

Tissue procurement and processing

Human sample acquisition and experimental procedures were carried out in compliance with regulations and protocols approved by the IRB-HSR at the University of Virginia in accordance with the U.S. Common Rule. In accordance with IRB Protocol #19272, breast cancer samples were collected as ultrasound-guided core needle biopsies during diagnostic visits from participants who provided informed consent. Each core biopsy was cut into two pieces, freshly cryoembedded in NEG-50 medium (Richard-Allan Scientific) in a dry ice-isopentane bath, and stored at −80°C wrapped in aluminum foil. Cryosectioning and slide storage was performed exactly as described previously (16).

Rapid histology–immunofluorescence and laser-capture microdissection

For samples with low immune infiltration (UVABC1, UVABC2, and UVABC4) slides were stained and dehydrated as described previously (20). For samples with high immune infiltration (UVABC3 and UVABC5), slides were fixed immediately in 75% ethanol for 30 seconds, rehydrated quickly with PBS, and stained with a mixture of Alexa 488-conjugated KRT8 antibody (Abcam ab192467, 1:20 dilution), YO-PRO-3 (Fisher/Invitrogen Y3607, 1:1000 dilution) and 1 U/ml RNAsin-Plus (Promega) in PBS for one minute. Slides were rinsed twice with PBS before dehydrating with 70% ethanol for 15 seconds, 95% ethanol for 15 seconds, and 100% ethanol for one minute and finally clearing with xylene for two minutes.

Slides were microdissected immediately on an Arcturus XT LCM instrument (Applied Biosystems) using Capsure HS caps (Arcturus). Cells were either visualized by brightfield microscopy (UVABC1, UVABC2, and UVABC4) or with a dual FITC/TRITC filter (UVABC3, UVABC5). The smallest spot size and typical instrument settings (~50 mW power and ~2 msec duration) yielded ~25 μm spot diameters capturing 1–3 breast carcinoma cells per laser shot.

For each biopsy, adjacent clusters of 10 cells were captured as 10-cell samples throughout multiple cryosections to access different regions of the tumor. In addition, a matched number of cells was captured nearby the 10-cell samples on the same LCM cap, extracted as a pool, and diluted into 10-cell equivalents that serve as measurement controls (pool-and-split controls). For each tumor, 10cRNA-seq datasets include 10-cell samples and matched pool-and-split controls captured across multiple days.

RNA extraction and library preparation

RNA extraction and amplification of microdissected samples was performed as described previously (16). Briefly, RNA was eluted from the LCM caps by digesting with proteinase K, and oligo-dT primed cDNA was synthesized. Residual RNA was degraded by RNAse H (NEB) digestion, and cDNA was poly(A) tailed with terminal transferase (Roche). Poly(A)-cDNA was amplified using an AL1 primer (ATTGGATCCAGGCCGCTCTGGACAAAATATGAATTCTTTTTTTTTTTTTTTTTTTTTTTT) and a blend of Taq polymerase (NEB) and Phusion (NEB) for 30 cycles.

Poly (A) PCR-amplified samples were first assessed by quantitative PCR for exogenous ERCC spike in standards and endogenous genes (*GAPDH* and *RPL8* as loading controls and the epithelial marker *KRT8*) as previously described (16). New primers for this study were *KRT8* (Fwd: GCCGTGGTTGTGAAGAAGAT, Rev: CCCCAGGTAGTAAACTCCCC) and *RPL8* (Fwd: CCCAGCTCAACATTGGCAAT, Rev: ACGGGTCTTCTTGGTCTCAG). Samples were retained if the geometric mean of quantification cycles for the *GAPDH–RPL8* loading controls was within 3x the interquartile range of the median calculated across all 10-cell samples of the biopsy. Samples beyond that range were excluded because of over- or under-capture during LCM. For samples with increased immune infiltration, we additionally excluded samples with a detectable quantification cycle for the T cell marker *CD3D* (Fwd: TGCTTTGCTGGACATGAGACT, Rev: CAGGTTCACTTGTTCCGAGC).

Libraries for sequencing were re-amplified, purified, and tagmented as described previously (16). Briefly, each poly(A) PCR cDNA sample was re-amplified for a number of cycles where the amplification remained in the exponential phase (typically 10 to 20). Re-amplified cDNA was then twice purified with Ampure Agencourt XP SPRI beads. After bead purification, samples were quantified on a CFX96 real-time PCR instrument (Bio-Rad) using a Qubit BR Assay Kit (Thermo Fisher). Samples were diluted to 0.2 ng/µl and tagmented with the Nextera XT DNA Library Preparation Kit (Illumina).

RNA sequencing

Libraries from 10-cell samples were multiplexed at an equimolar ratio, and 1.3 pM of the multiplexed pool was sequenced on a NextSeq 500 instrument with NextSeq 500/550 Mid/High Output v1/v2/v2.5 kits (Illumina) to obtain 75-bp paired-end reads. From the sequencing reads, adapters were trimmed using fastq-mcf in the EAutils package (version ea-utils.1.1.2-779), and with the following options: -q 10 -t 0.01 -k 0 (quality threshold 10, 0.01% occurrence frequency, no nucleotide skew causing cycle removal). Quality checks were performed using FastQC (version 0.11.8) and MultiQC (version 1.7). Data were aligned to the human transcriptome

(GRCh38.84) along with reference sequences for ERCC spike-ins using RSEM (version 1.3.0) and Bowtie 2 (version 2.3.4.3). RSEM options for the 10cRNA-seq data also included the following: --single-cell-prior --paired-end. RSEM read counts were converted to transcripts per million (TPM) by dividing each value by the total read count for each sample and multiplying by 10^6. Mitochondrial genes and ERCC spike-ins were not counted towards the total read count during TPM normalization.

Molecular subtype assignments

Microarray-based PAM50 centroids and associated code were obtained from the UNC Microarray Database (https://genome.unc.edu/pubsup/breastGEO/PAM50.zip). To adapt the signature for RNA-seq data, RSEM aligned TPM data for TCGA breast tumors was obtained from the UCSC genome browser (26). Using a balanced number of estrogen receptor (ER)-negative and ER-positive tumors, median RNA-seq values for PAM50 genes were calculated and subtracted from the entire cohort for standardization. Standardized values were used to predict PAM50 subtypes by using downloaded centroids and code from the UNC Microarray Database (27). Successful model training was visualized by a principal component plot showing both training (microarray) and test (TCGA RNA-seq) data clustering by molecular subtype. The same median correction method was attempted for 10cRNA-seq data from UVABC tumors, but model calibration was unsuccessful due to a lack of ER-negative samples and some large differences in overall abundance of PAM50 genes between bulk and 10-cell data. As a substitute, 10cRNA-seq samples were scored for transcriptional modules associated with ESR1 (464 genes), ERBB2 (27 genes), and AURKA (229 genes) using the “subtype.cluster” function within the package “genefu” (version 2.16.0) in R. On the basis of these module scores, samples were subtyped as Luminal A (ESR1+, ERBB2–, AURKA–), Luminal B (ESR1+, ERBB2–, AURKA+), HER2 (ERBB2+), and Basal (ESR1–, ERBB2–).

UMAP projections

All UMAP projections were generated using the R package “umapr” (version 0.0.0.9001) with the following parameters: neighbors = 4, distance metric = "correlation", minimum distance = 0. For UVABC embedding, RSEM counts for UVABC tumors were converted to TPM values and projected onto a UMAP using all endogenous transcripts. The same UMAP projection was re-colored by subtype classifications and batch number. Batch effects were excluded by visualizing TPM estimates of exogenous ERCC spike-in expression for all UVABC tumors on a separate UMAP. scRNA-seq data of breast tumors (28) were obtained as RSEM-aligned TPM

values from the Gene Expression Omnibus (GSE75688). Epithelial carcinoma cells were separated from infiltrating cell types in the scRNA-seq data by selecting cells that expressed *KRT8* and *EPCAM* at TPM > 1. 10cRNA-seq UVABC samples and filtered scRNA-seq data (78 cells) were merged by transcript and projected on the same UMAP.

To account for tumor-specific differences in overall transcript abundance, we centered the expression of every transcript by the 25th quartile of its expression in 10-cell samples from each tumor. Quartile-centered samples for each tumor were then merged by transcript and projected onto a UMAP. From the merged quartile-centered data, expression of 710 RHEGs was extracted and projected onto a separate UMAP.

Overdispersion-based stochastic profiling

10cRNA-seq analysis consisted of an identical algorithm applied separately to 28 10-cell samples and 20 pool-and-split controls from each UVABC tumor. RSEM values were rounded to integer counts, and transcripts with zero counts throughout were removed. Abundance-dependent expected expression and error models were generated separately for 10-cell samples and pool-and-split controls using the "knn.error.models" function in the package "scde" (version 1.99.4) with *k* nearest neighbors set to ¼ of the sample size for both sets. Only transcripts that had a minimum transcript count of 5 (min.count.threshold) in at least 5 samples (min.non.failed) were considered for model generation. From the abundance-adjusted error models, adjusted-variance estimates of overdispersion were calculated using the "pagoda.varnorm" function of the same package using the generated error models as input. The variance was further adjusted to account for read-depth using the "pagoda.subtract.aspect" function. Transcripts with adjusted variances in 10-cell samples that exceeded the 95th percentile of adjusted variances in pool-and-split controls were considered candidate heterogeneously expressed transcripts. Finally, transcripts with adjusted variances in the top 5th percentile of pool-and-split controls reflecting high measurement variability were filtered out of candidate lists. Overdispersion-based stochastic profiling was applied identically to each of the five UVABC cases to obtain candidate gene lists. Final candidate gene lists for every tumor are supplied along with RHEGs in Supplementary File S1.

Cohort subsampling for RHEG estimation

The five UVABC tumors were exhaustively downsampled as groups of n = 1 (five total possibilities), 2 (10 total), 3 (10 total), 4 (five total), or 5 (one total) and intersected with the

operational RHEG definition of transcript heterogeneities that recur in ≥50% of the cases considered: one for n = 1 or 2, two for n = 3 or 4, and three for n = 5.

CNV inference

To predict CNVs, we used InferCNV, which corrects the input expression data (here, 10cRNA-seq) for average gene expression based on normal reference cells and applies a moving average with a sliding window of 101 genes within each chromosome. Smoothed expression values are once again corrected against the normal reference and estimated copy number alterations are reported. Normal breast tissue gene-expression data was obtained from GTEx (29) as a reference dataset for copy-number variation. The reference GTEx data, UVABC 10cRNA-seq data, and a reference genome position file (GRCh38.86) were input to the "CreateInfercnvObject" function in the package "InferCNV" (version 1.0.3) in R (30). The InferCNV object was then analyzed with the "infercnv::run" function with dynamic threshold denoising to infer copy-number variations as previously described (31).

Periodicity of cell-cycle RHEGs vs. non-RHEGs

We obtained 361 oscillating transcripts from Cyclebase 3.0 (32), reconciled aliases with official gene names, and intersected with the RHEGs from 10cRNA-seq, obtaining 24 shared transcripts. Next, using microarray data from synchronized HeLa cells (33), we identified probesets for 21 of the 24 shared transcripts along with those of the top 10 cycling transcripts according to Cyclebase 3.0 (32). The centered probeset data and HeLa pseudotime estimates are available through Cyclebase 3.0 for three complete cell cycles, but only the first two cycles show strong synchronization (32). We calculated the time interval above the mean value and compared it as a ratio to each of the adjacent time intervals below the mean value. For the ratio, the larger pseudotime interval (above or below the mean) was placed in the numerator. We calculated the skewness of the ratio distributions (indicating time-interval asymmetry above-vs.-below the mean) for the 21 RHEGs with identifiable probesets vs. the top 10 cycling transcripts and estimated uncertainty by bootstrapping.

Monte Carlo simulations of three-state stochastic profiling

Monte Carlo simulations of stochastic profiling under the assumption of two regulatory states were performed in MATLAB as described with available software (19). The two-state model assumes a binomial distribution for the cellular dichotomy and log-normal distribution of measured transcripts. To build a three-state model reflecting cell cycle-regulated variation, we

used a multinomial distribution to reflect cell-cycle fractions, two lognormal regulatory states for G1 and G2/M phases, and a uniform "S-phase" interval between the two other regulatory states. Two- or three-state distributions were compared against a null distribution of lognormal variation using the Kolmogorov-Smirnov test. The simulations for a parameter set were run 50 times to measure the median *p* values and the associated nonparametric confidence intervals. Stochastic sampling was deemed effective when the median *p* value for $F \neq 0$ (multiple states) was less than 0.05 and the median *p* value for $F = 0$ (one state) was greater than 0.05. MATLAB code for the modified three-state distributions is available in Supplementary File S2.

Gene signature overlaps with RHEGs

All overlaps were viewed using the "venn" function in the R package gplots (version 3.0.1.1), and intersections were obtained using the "intersect" function. Significance of overlap was calculated using the hypergeometric test in R through the "phyper" function and a total gene count of 20,000. For assessing detachment-induced artefacts in the RHEG gene set, we obtained a murine detachment-induced gene signature (14) and converted mouse genes to human orthologs with the Ensembl biomart in R. Any remaining mouse gene names were capitalized in accordance with human gene symbol conventions. The human ortholog mapping was verified against the human transcriptome reference gene list used for 10cRNA-seq data. Breast driver genes were obtained from a larger gene list of cancer drivers (23) filtered for tissue of origin. We similarly assessed overlap with transcripts altered by CTCF knockout in luminal breast cancer cells (34), GATA3 target genes (35), EMT signature gene sets (24,36), and an aggregated pan-cancer driver gene set (23,37). For the last comparison, nine RHEGs were considered proximal to driver genes (Supplementary Table ST2) and treated as equivalent between the two sets. All mismatched gene aliases were corrected before overlap testing.

Statistics

Sample sizes for stochastic profiling were determined by Monte Carlo simulation (19). Differences in genes detected per sample between 10cRNA-seq and scRNA-seq were assessed by Kolmogorov-Smirnov test using the "ks.test" function in R. Significance of overlaps between candidate genes identified in different UVABC tumors were assessed by Monte Carlo simulations that drew the total number transcripts per tumor randomly from a common pool of 14,824 genes (total transcripts eligible for overdispersion analysis in all tumors). Observed overlaps were compared with 1000 Monte Carlo simulations to estimate a *p* value, which was adjusted for multiple comparisons by using the Šidák correction. All overlaps between the

RHEG set and other gene sets were assessed for significance by hypergeometric test using the "phyper" function in R and a background of 20,000 genes. Kolmogorov-Smirnov tests for Monte Carlo simulations for stochastic profiling were assessed using the "kstest" function in MATLAB. Hierarchical clustering was performed using "pheatmap" in R using Euclidean distance and "ward.D2" linkage.

Data availability

10cRNA-seq data from this study is available through the NCBI Gene Expression Omnibus (GSE147356, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147356 Reviewer token: wzcfouoijfodzij).

Results

Carcinoma-focused 10-cell profiling of early-stage luminal breast cancer

Women were selected for enrollment if they required ultrasound-guided biopsy for a suspected malignancy after screening mammography (BI-RADS 4C and higher). Just before diagnostic biopsy, we obtained written informed consent to collect an additional ultrasound-guided core sample, which was cryoembedded (16) within one minute of acquisition (**Fig. 1A** and **Table 1**). After clinical diagnosis of hormone-positive, *HER2*-negative breast cancer, we selected five cases that were as closely matched as possible (Supplementary Table ST1). The median tumor biopsy was Stage 1, Grade 3, and aged 63 years—the late-onset group for breast cancer (38).

Despite gross similarities in tumor characteristics, we noted elevated lymphocyte infiltration in two cases (UVABC3 and UVABC5) that rendered them problematic to microdissect by nuclear histology alone (**Fig. 1B** and Supplementary Table ST1). Therefore, we devised an immuno-LCM (39) procedure that combines an Alexa Fluor 488-conjugated, high-affinity monoclonal antibody against KRT8 with the red-orange nucleic acid stain YOPRO3 (see Materials and Methods). A one-minute incubation with the antibody-dye cocktail was sufficient to resolve KRT8-positive carcinoma cells from KRT8-negative stromal cells and the YOPRO3-negative autofluorescence of tissue collagen (Supplementary Fig. S1A). We could not find any evidence that the antibody or dye interfered with the critical early steps of 10cRNA-seq (Supplementary Fig. S1B) (16).

For each case, we collected 28 random pools of 10 carcinoma cells located throughout cryosections of the core biopsy. Pools were assembled as local 10-cell groups that reflect both clonal and microenvironmental heterogeneity. We recorded the spatial position of all cells microdissected in the 10-cell samples to leave open the possibility of retroactively linking transcriptomic changes to histological or topological features of the tumor. Samples were deeply sequenced at 6.04 ± 0.75 million reads per 10-cell pool to ensure saturation of gene detection and provide maximum sensitivity for identifying non-carcinoma contaminants. Across all cases, we found that the luminal markers *KRT8* and *ESR1* predominated by transcripts per million (TPM), whereas markers for myoepithelial cells (*KRT14*, *KRT5*), T cells (*CD3D*), B cells (*CD19*), and macrophages (*FGCR1A*) were rarely detected (**Fig. 1C**). Even though desmoplasia was marked for all but one case (Supplementary Table ST1), fibroblast contamination was problematic in only ~14% of 10-cell samples (using $\log_2$ *FAP* > 5 transcripts per million [TPM] ≈ 0.8 copies per cell (16) as a stringent threshold). These samples were

excluded retroactively for downstream analyses related to EMT signatures (see below). Overall, the observations confirmed the fidelity of (immuno-)LCM for isolating spatially-resolved, carcinoma-specific transcriptomic profiles with minimal disruption of tumor architecture.

10cRNA-seq transcriptomes retain the inter-tumor and intra-tumor heterogeneity of luminal breast carcinoma cells profiled by scRNA-seq

As a first assessment, the 10cRNA-seq data were visualized by uniform manifold approximation and projection (UMAP) (40). Consistent with past descriptions of carcinoma heterogeneity by scRNA-seq (28,41,42), we found that 10cRNA-seq data clustered tightly by patient (**Fig. 2A**). Batch effects were not evident within the patient clusters (Supplementary Fig. S2A–E). Furthermore, we did not observe any clustering in a separate UMAP visualization using only the data from ERCC spike-ins added to every sample at the time of RNA extraction (Supplementary Fig. S2F). The observed separation of 10cRNA-seq transcriptomes (**Fig. 2A**) thus reflects bona fide inter-tumor differences between cases.

Next, we sought to classify individual 10cRNA-seq samples into intrinsic breast cancer subtypes (43). We adapted the microarray-based PAM50 classification of subtypes (27) to RNA-seq (26), but the 50-gene signature was not robust enough for 10cRNA-seq observations (Supplementary Fig. S3). As a substitute, we used expression-signature modules (44) associated with *ESR1*, *ERBB2*, and *AURKA* as proxies for the hormone, *HER2*, and proliferative status of each 10-cell sample (see Materials and Methods). Within patients, there was considerable variability in module scores (**Fig. 2B**), corroborating an earlier scRNA-seq study of multiple breast-cancer subtypes (28). Although most 10-cell samples were classified as luminal A or luminal B subtype, all cases but UVABC4 contained observations scoring more strongly to other subtypes (**Fig. 2C**). Two cases (UVABC1 and UVABC3) harbored instances of all four subtypes, analogous to scRNA-seq observations in glioblastoma (31). The clustering of mixed classifications implied that other patient-specific gene programs were dominant in the UMAP organization. Variations in subtype class were repeatedly documented in nearby samples microdissected from the same histologic section (Supplementary Fig. S4). Even for early-stage breast tumors, the 10cRNA-seq data suggested that local variations in regulatory state are pervasive.

Toward a more-direct comparison of 10-cell data with single-cell measurements of gene expression, we extracted 78 *KRT8*$^{+}$*EPCAM*$^{+}$ carcinoma cells from three cases of luminal breast cancer profiled by scRNA-seq (28) (see Materials and Methods). As previously reported by us and others (16,17,45,46), there were significantly more transcripts detected in local 10-cell pools

compared to singly isolated cells (10,066 ± 1,416 genes vs. 5,957 ± 1,824 genes; $p < 10^{-15}$ by K-S test; **Fig. 2D**). Overall, 3319 transcripts found in 10cRNA-seq pools were entirely undetected by scRNA-seq. Notwithstanding the differences in gene coverage, when scRNA-seq and 10cRNA-seq samples were projected on a shared UMAP, the separation between methods was comparable to that among patients (see Materials and Methods; **Fig. 2E**). Together, the data argue that 10-cell pooling does not dilute out the cell-to-cell and tumor-to-tumor heterogeneities recognized by scRNA-seq.

Stochastic profiling by 10cRNA-seq identifies candidate regulatory heterogeneities

To go beyond qualitative descriptions of molecular subtype and inter-tumor differences, we sought to adapt the theory of stochastic profiling (19,20) to 10cRNA-seq (16). Fluctuation analysis by RNA-seq brings additional challenges and opportunities compared to microarray-based transcriptomics. The data are not biased by the position or quality of hybridization probes [as discussed in (16)], but they are sensitive to read depth, and low-abundance transcripts are susceptible to noise from counting statistics. Specifically, the discrete and left-censored character of rare transcripts partially obscures sample-to-sample fluctuations (**Fig. 3A**) and deviates from lognormally distributed models used in earlier analyses (19,20,22). We surmounted these hurdles by extracting an abundance-dependent dispersion module from the SCDE package (47,48) and redeploying it as a separate inference tool for stochastic profiling. The module relates the squared coefficient of variation (CV) of each gene in a study (here, a patient) to the abundance magnitude of that gene, building an expectation model of variance at a given abundance (**Fig. 3B**). The variance of each transcript is then normalized by the expected variance for that transcript's abundance, yielding an overdispersion score for the transcript. For high-abundance genes, overdisperse transcripts show multiple modes or heavier tails than expected (**Fig. 3C** and **3D**). Low-abundance genes with overdispersion are skewed by multiple instances of moderate-to-high TPM (**Fig. 3E** and **3F**). The dispersion module incorporates discrete negative-binomial and Poisson processes to model aligned reads and dropouts. The overdispersion score thus provides a principled metric for stochastic-profiling analysis of 10cRNA-seq data.

In conventional scRNA-seq, each cell is considered as an N-of-1 observation that convolves biological variability and technical noise. For our study, technical noise could be quantified more rigorously by pool-and-split sequencing of 10-cell equivalents from hundreds of carcinoma cells microdissected in the same vicinity as the samples. Accordingly, we sequenced 20 pool-and-split controls in parallel with the 28 10-cell samples, analyzing the

controls separately to construct a null distribution for transcript overdispersion in each tumor. The upper 95th percentile in the null model defined an overdispersion cutoff for the 10-cell samples—transcripts above this threshold in the 10-cell samples (but not in the null) were considered candidate regulatory heterogeneities (**Fig. 3G**, **3H**, and Supplementary Fig. S5).

Stochastic profiling identifies recurrent transcriptional regulatory heterogeneities

By abundance-dependent dispersion, stochastic profiling identified 9206 candidate heterogeneities in the UVABC cohort, 161 transcripts of which were undetected by scRNA-seq (28). Only a few percent of candidate genes were found in loci of inferred copy-number variation (Supplementary Fig. S6), excluding major contributions from heritable differences among subclones in a tumor. 3627 candidates were exclusive to one of five patients, laying bare the extraordinary challenge of interpreting malignant cell-state heterogeneity beyond well-known markers of differentiation (12). Encouragingly, when the candidates were intersected, we noted a significant enrichment of transcripts that appeared repeatedly in three or more breast-cancer cases (**Fig. 4A**). Such transcripts were classified as recurrent heterogeneously expressed genes (RHEGs), for which there were 710 in total. Generalizing the RHEG definition to candidates appearing in >50% of the cases considered, we examined the stability of RHEG numbers by subsampling the UVABC cases (see Materials and Methods). As the quantity of patients increased, RHEGs stabilized in the range of 500–1500 for this highly circumscribed cohort (**Fig. 4B** and Supplementary Table ST1). RHEGs provide a conceptual framework for prioritizing cell-state regulatory heterogeneities identified in vivo (49,50).

To evaluate RHEGs as an organizing principle, we revisited the UMAP visualization of UVABC cases from the standpoint of regulatory heterogeneity (**Fig. 2A**). Because abundance-dependent dispersion evaluates fluctuations local to each tumor (**Fig. 3**), it was first necessary to standardize the 10cRNA-seq transcriptomes separately and regenerate the UMAP (see Materials and Methods). Tumor-specific standardization intermingled the 10cRNA-seq observations considerably, but two clusters remained enriched in samples from UVABC1 and UVABC2 (**Fig. 4C**). When the same approach was applied using RHEGs exclusively, we observed a projection that was different from when the whole 10cRNA-seq transcriptome was used (**Fig. 4D**). For the same UMAP parameters, samples were more distributed than clustered, with the UVABC1-enriched cluster disappearing and a new UVABC4-enriched cluster appearing. This analysis suggested that RHEGs could be used as a lens to refocus transcriptome-wide heterogeneity on the most-robust variations.

RHEGs are not dominated by cell-cycle covariates

In scRNA-seq, the most-overarching contributor to heterogeneity is the phase of the cell cycle (41,51,52). However, it was not obvious whether such single-cell variations would also overrepresent in RHEGs derived from 10-cell pools. Using a panel of 863 transcripts associated with replicating cells (51), we identified 62 among the 710 RHEGs, a significant overlap ($p < 10^{-6}$ by hypergeometric test) but one comprising <10% of the list overall. Most of the overlapping genes were expressed acutely during one cell-cycle transition (e.g., G1/S, G2/M), which is akin to the two-state expression models foundational for the theory of stochastic profiling (19,20,22). When the cell-cycle search was restricted to 361 oscillating transcripts (32), the RHEG intersection reduced to 24 genes ($p < 0.01$ by hypergeometric test). Moreover, when we compared the periodicity of the overlapping genes to the most-symmetrically cycling transcripts, we found that RHEGs were significantly more biased toward up- or downregulation (see Materials and Methods; Supplementary Fig. S7A and S7B and Supplementary File S3). These results are consistent with Monte-Carlo simulations of stochastic profiling that model a three-state distribution corresponding to G1, S, and G2/M populations (Supplementary Fig. S7C–F). Stochastic-profiling theory thus bolstered our experimental results indicating that cycling transcripts contribute <5% to the RHEGs identified.

RHEGs are largely devoid of detachment artifacts and influence from breast-cancer driver genes

We also considered other trivial explanations for genes in the RHEG set. Although tumors were not dissociated, it was possible that detachment-like regulatory variation was induced locally and rapidly from the tissue damage of the biopsy procedure. Using a 138-gene signature for dissociation-induced transcripts in muscle satellite cells (14), we intersected with the RHEG set and found that only two were shared (**Fig. 5A**). This marginal under-enrichment ($1 - p < 0.05$) indicated that our clinical-procurement and sample-handling procedures had avoided detachment-like damage responses in the breast carcinoma cells.

Next, we looked at known breast cancer drivers with the premise that mutations may arise subclonally in a breast carcinoma (53) and disrupt abundance of the encoded transcript (54). Among 29 robust driver genes for breast cancer (23), only one was shared with the RHEG set: *CTCF* (**Fig. 5B**, left). As an insulator protein, CTCF abundance changes could cause secondary transcriptional alterations; however, we did not observe any enrichment for conserved CTCF-sensitive genes (34) in the RHEG set (**Fig. 5B**, right). Although not classified as a RHEG, we also investigated transcriptional targets for the most-prevalent transcription

factor mutated in luminal breast cancer, GATA3 (23). Again, we found no RHEG enrichment among 1213 transcripts altered by mutant GATA3 in luminal breast cancer cells (35) (**Fig. 5C**). The lack of association collectively supported that RHEGs were more than a reflection of known sources of cell-to-cell heterogeneity in cancer.

RHEGs are enriched for EMT signatures and correlate with canonical EMT markers

scRNA-seq of dissociated tumors has identified (partial-)EMT states in some carcinomas (41,42) but not others (55). EMT-like transcriptional profiles also arise normally in the cell type of origin for serous ovarian cancer (56). For breast cancer, changes along the EMT spectrum are mostly described in hormone-negative cell lines, but more-recent work reports EMT-like activation patterns in 65–85% of primary luminal breast cancers (57). The literature thus supported a focused search for EMT states among RHEGs.

We intersected a pan-cancer EMT signature (36) with the RHEG set and found significant overlap in multiple collagens, matricellular proteins, and other transcripts in the signature (**Fig. 6A**). The data suggest that ECM dysregulation in these tumors is jointly mediated by the carcinoma cells together with cancer-associated fibroblasts. RHEG enrichment was also found with an independently derived EMT signature (24) (Supplementary Fig. S8A), reinforcing the result. Formally, none of the canonical EMT regulators [*ZEB1* (M), *ZEB2* (M)] or markers [*CDH1* (epithelial, E), *VIM* (mesenchymal, M), *FN1* (M)] were RHEGs, even though all were detected reliably enough for stochastic-profiling analysis. We clustered these canonical EMT transcripts with the EMT RHEGs after stringently removing samples with any evidence of fibroblast contamination (14% of samples with *FAP* > 5 TPM ≈ 8 copies in one cell of the 10-cell pool). There was clear separation of E- and M-associated transcripts among 10-cell pools along with several notable subclusters by gene and by sample (**Fig. 6B**). The organization by patient was unexpected; for example, UVABC2 showed the most evidence for the E state, even though it was one of the most-advanced stage tumors of the cohort (Supplementary Table ST1). Reciprocally, 10-cell profiles of the high-grade UVABC3 tumor were no more scattershot in EMT transcripts than UVABC4 (grade 2) or UVABC5 (grade 1) (**Fig. 6B**). Among samples with M characteristics, *ZEB2* appeared to track with those samples abundant for some transcripts (*FN1*, *COL6A1*, *SPARC, VIM*) but not others (*COL5A2*, TAGLN). M-state fragmentation was also observed in UVABC1, which was predominated by samples positive for *VIM*, *SPARC*, and *SERPING1* but negative for CTSK, TAGLN, and mesenchymal collagens (**Fig. 6B** and Supplementary Fig. S8B). The RHEG set thereby provided a transcriptomic resource for looking more deeply at EMT regulatory patterns in early-stage luminal breast cancers.

RHEGs are enriched for pan-cancer driver genes and suggest transcription factor-target relationships in single cells

The dearth of breast-cancer driver genes among RHEGs (**Fig. 5B**) prompted us to look at cancer drivers more broadly. We merged 299 robust drivers for any cancer type (23) with the latest pan-cancer analysis reporting 803 drivers from 2,658 tumors (37) and intersected with the RHEG set. There were multiple instances where RHEGs resided in the same complex, pathway, or gene subfamily as a cancer driver (Supplementary Table ST2). We included these proximal RHEGs and altogether found 46 genes as "RHEG drivers" shared between the two datasets (p = 0.001 by hypergeometric test; **Fig. 6C**). Even with the expanded driver set, we found no enrichment for mutated breast-cancer driver genes (p = 0.6 by hypergeometric test). In the UVABC cohort, RHEG drivers may be leveraged noncanonically through transcriptional regulation rather than mutation.

Last, we clustered the RHEG drivers to ask whether there were interpretable covariations spanning multiple patients (**Fig. 6D**). Associations among 10-cell samples were tightest for UVABC2 and UVABC4, in line with their separation on the earlier UMAP (**Fig. 4D**). Repeatedly, co-clustering RHEG drivers suggested direct modes of action between transcription factors and target genes (**Fig. 6D**, arrows). For example, knockdown of *NFATC4* blocks induction of the neighboring RHEG driver, *TNFSF10* (58), and there is literature that the reprogramming factor *KLF4* is required for full induction of *CDKN1A* (59). Although no functional studies are available for *CDKN2D*, another co-clustering cyclin-dependent kinase inhibitor (**Fig. 6D**, arrows), the *CDKN2D* locus is occupied by KLF4 (60) and may warrant further study. Likewise, the *MAD1L1* locus is reportedly among the top 250 binding events in the genome for *TP73* (61)—a RHEG driver absent from all luminal breast cancer cells profiled by scRNA-seq (28). Some of the debate involving *MAD1L1* as a TP53 target gene [reviewed in (62)] could be explained by compensation from TP73 (63). RHEG drivers are variably expressed within luminal breast cancers, and our data suggest that some are variably active.

Discussion

This work combines 10cRNA-seq (16) and stochastic profiling (20) for disruption-free isolation of cancer cell-regulatory heterogeneities in a clinically practicable way. We targeted LCM isolation to breast-carcinoma cells here by using nuclear cytology or epithelial-targeted antibodies; the approach is also compatible with genetically encoded fluorophores (49,50). For 10cRNA-seq, artifactual cell stress (14) is avoided by LCM (49), and dominant cycling transcripts are mitigated by 10-cell averaging. But in many respects, 10cRNA-seq of malignant cells shares similarities with scRNA-seq: cases are very different from one another, and samples vary substantially within cases. What differs is overall gene coverage per sample (10cRNA-seq > scRNA-seq), as well as the analytical approach needed to discern single-cell differences. Abundance-dependent overdispersion can identify candidates from 10cRNA-seq, much like the nonparametric distribution tests first deployed for microarray data (19). As with microarrays, we anticipate future developments toward parameterizing the underlying single-cell distributions, which combine to yield 10cRNA-seq observations (22). The 10cRNA-seq-based subtype classifications predicted local differences not obvious from histology, and tools for spatial analysis of biomolecules are rapidly advancing (64,65). It will be especially intriguing when spatial transcriptomics (66) reaches the resolution and sensitivity of 10 cells.

RHEGs open the possibility of making more specific claims about intratumor heterogeneity beyond cell stress, cell cycle, and cell type (12). Partial EMTs in carcinomas have been documented by scRNA-seq (28,41), which we verify here in earlier-stage tumors without any pre-dissociation (**Table 1**). While there are many ways to elicit EMT-like states, a leading explanation for the UVABC cohort is tissue stiffness (67) given their marked desmoplasia. Notably, one RHEG is the hemidesmosomal integrin *ITGB4*, which acts as a critical sensor for matrix stiffness in breast epithelia (5). *ITGB4* was undetected in every luminal breast cancer cell analyzed by conventional scRNA-seq (28).

Secreted ligands and receptors (68) are not overly prevalent among RHEG drivers, but a pair with some coordination is the p53 target gene *GDF15* (69) and its cognate receptor *RET* (70) (**Fig. 6D**). In an accompanying study on another carcinoma type (49), we suggest that such receptor-ligand pairs could engage as locally varying autocrine–paracrine circuits within a tumor and shape the immune microenvironment. For instance, among the 136 candidates shared by the two cases requiring immuno-LCM because of extensive lymphocytic infiltration, we identified an inhibitory ligand for NK cells [*ADGRG1* (71)], a major histocompatibility class II receptor [*HLA-DRA* (72)], a macrophage stimulatory ligand [*MST1* (73)], and the

palmitoyltransferase for PD-L1 [*ZDHHC9* (74)]. Natural variation in such carcinoma transcripts could one day be mapped to associating changes in the type and extent of immune-cell recruitment (75).

Single-cell cancer biology must trade off coverage, throughput, and handling artifacts to retain the conceptual allure of measuring one cell. The approaches described and implemented here for late-onset, early-stage breast cancer are also a compromise, but one that triangulates differently by using cell pools to reduce handling and improve coverage. We see great potential in using stochastic profiling by 10cRNA-seq to deconstruct the very earliest stages of tumor initiation and premalignancy in engineered systems (50) and in precancerous *in situ* lesions of the breast where the need for treatment is actively debated (76).

Figures

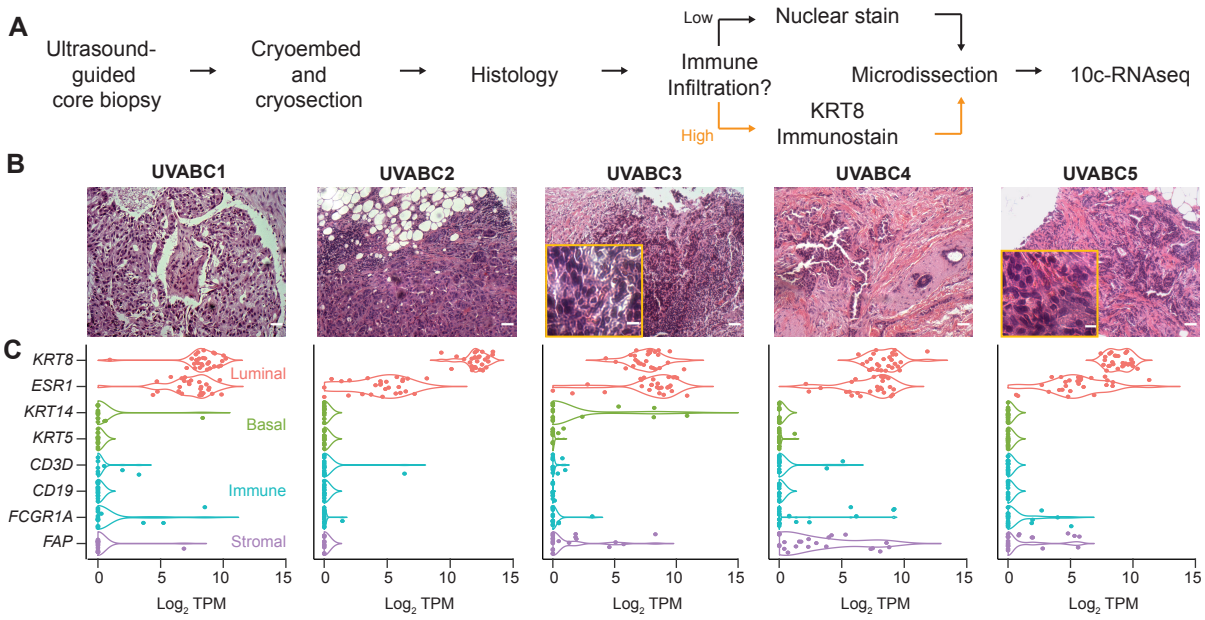


Figure 1. Focused transcriptional profiling of breast carcinoma cells without dissociation in early-stage tumor biopsies. **A,** Biopsy tissue was immediately cryoembedded and later cryosectioned for laser-capture microdissection. Tumor cells were visualized with either a rapid nuclear stain or KRT8 immunostain (Supplementary Fig. S1) before 10cRNA-seq (16). **B,** Tumor histology of the UVABC cohort visualized by hematoxylin-eosin staining. Two cases (marigold inset) showed increased tumor infiltrating lymphocytes requiring KRT8 immunostain-guided LCM. **C,** Selective capture of epithelial carcinoma cells assessed by marker transcripts for luminal (*KRT8*, *ESR1*), basal (*KRT14*, *KRT5*), immune (*CD3D*, *CD19*, *FCGR1A*), and stromal (*FAP*) cells. Scale bar is 80 µm (**B**) or 10 µm (**B**, inset)

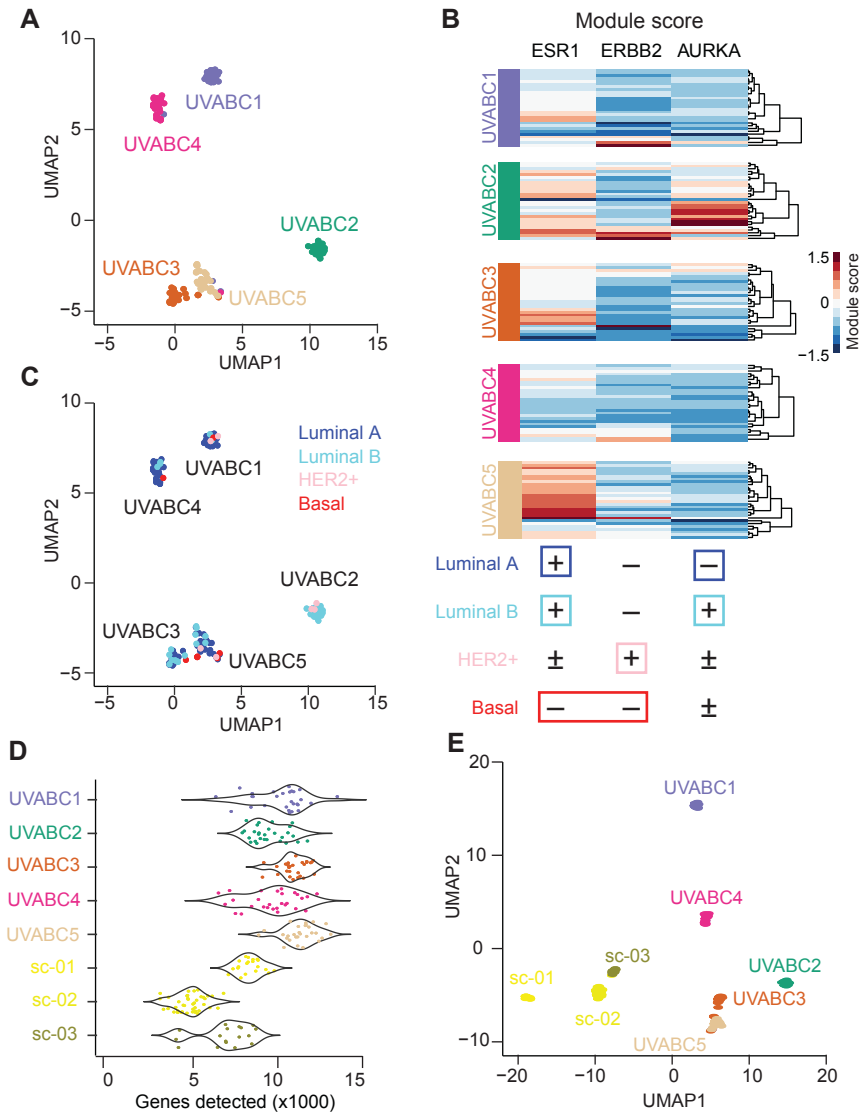


Figure 2. 10-cell transcriptomes of luminal breast carcinomas are heterogeneous among and within tumors. **A,** UMAP embedding of 10cRNA-seq samples from the UVABC cohort colored by tumor. **B,** A three-signature classification system for identifying molecular subtypes of breast cancer in 10cRNA-seq data. Module scores were used to classify samples as indicated. **C,** Molecular subtype classifications of the 10cRNA-seq samples projected as in **A**. **D,** Genes detected by 10cRNA-seq in the UVABC cohort compared to three luminal tumors profiled by scRNA-seq (sc-01, sc-02, sc-03) (28). $p < 10^{-15}$ by K-S test. **E,** UMAP embedding of tumors profiled by 10cRNA-seq and scRNA-seq.

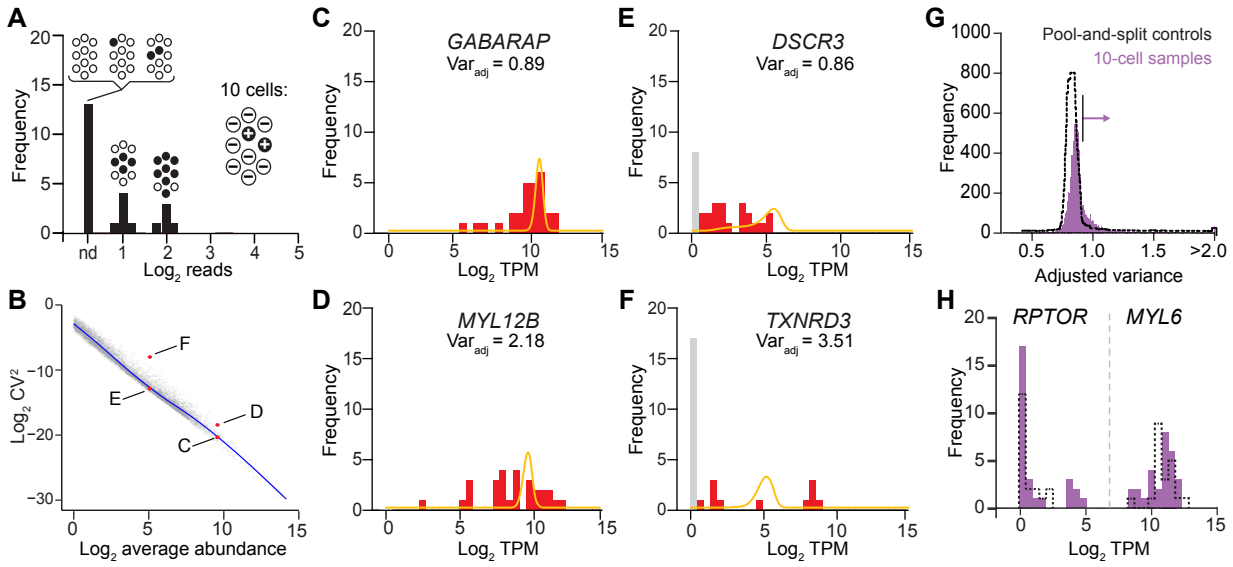


Figure 3. Stochastic profiling by 10cRNA-seq through abundance-dependent overdispersion statistics. **A,** Illustration of theoretical left-censored transcript, which is heterogeneously expressed at one copy per cell (filled, +) and measured with 50% efficiency. Low-frequency mixtures will be identically not detected (nd). **B,** Dispersion-abundance plot illustrating the expected inverse relationship. Example transcripts with similar abundance but different dispersion (red) are annotated by the **Fig. 3** subpanel in which they appear. The blue trace indicates a smoothing cubic spline fit of the summarized 10cRNA-seq fluctuations per transcript (gray). **C** and **D,** 10cRNA-seq sampling fluctuations for high-abundance transcripts with expected dispersion (*GABARAP*, **C**) and overdispersion (*MYL12B*, **D**). **E** and **F,** 10cRNA-seq sampling fluctuations for low-abundance transcripts with expected dispersion (*DSCR3*, **E**) and overdispersion (*TXNRD3*, **F**). **G,** Distribution of adjusted variance scores for each gene measured transcriptomically in separate 10-cell samples (purple) compared to pool-and-split controls estimating technical variation (black dashed). The arrow indicates the upper 5th percentile of adjusted variance for the pool-and-split controls, which is used as the cutoff for 10-cell samples. **H,** Examples of low-abundance (*RPTOR*) and high-abundance (*MYL6*) transcripts deemed to be significantly overdisperse given their relative abundance in 10-cell samples (purple) and their technical variation in pool-and-split controls (black dashed). For **C–F**, continuous traces (marigold) indicate the idealized dispersion expected given the observed transcript abundance, and the adjusted variance (Var_{adj}) is reported. Gray bars represent drop-out events, which are modeled by a separate posterior (47,48). For **B–H**, data from UVABC4 were used as representative examples.

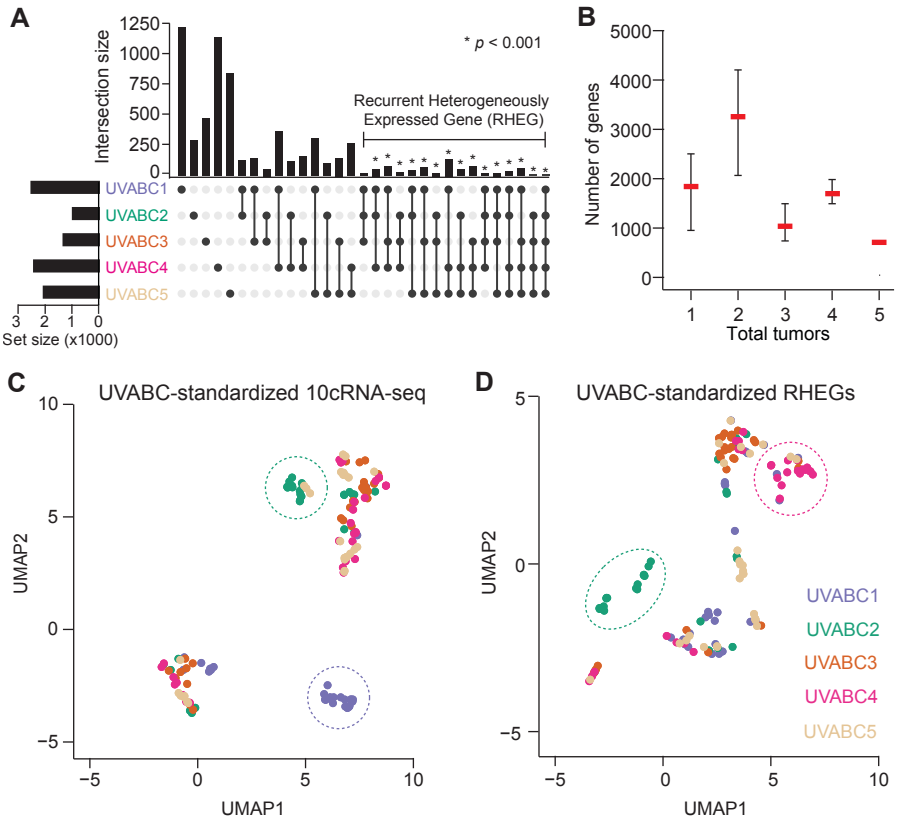


Figure 4. Intersecting candidate regulatory heterogeneities across tumors yields a set of recurrent heterogeneously expressed genes (RHEGs). **A,** Gene overlaps between and among cases in the UVABC cohort. Transcripts belonging to each tumor are available in Supplementary File S1. Enriched categories were assessed statistically by Monte-Carlo simulation (see Materials and Methods). **B,** RHEG size as a function of the number of UVABC tumors included. RHEGs were defined as transcripts that appear in >50% of the tumors included. Data are shown as the median ± range of ${}_5C_n$ combinations of *n* tumors in the UVABC cohort. **C** and **D,** UMAP embedding of patient-standardized 10cRNA-seq transcriptomes (**C**) or RHEGs (**D**). Patient-enriched clusters are highlighted with dashed ovals.

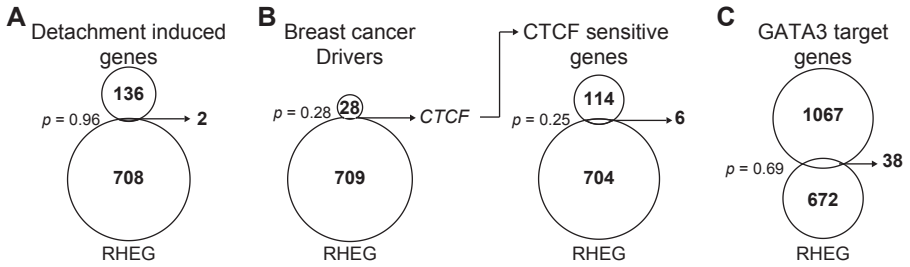


Figure 5. RHEGs have little in common with detachment signatures or mutational drivers of breast cancer. **A–C,** Venn diagrams intersecting the UVABC RHEG set with a cell-detachment signature (14) (**A**), a set of robust drivers for breast cancer (23) (**B**, left), a list of transcripts altered by CTCF knockout in a luminal breast cancer cell line (34) (**B**, right), and a list of GATA3 target genes (35) (**C**). Statistical significance of overlaps was assessed by the hypergeometric test.

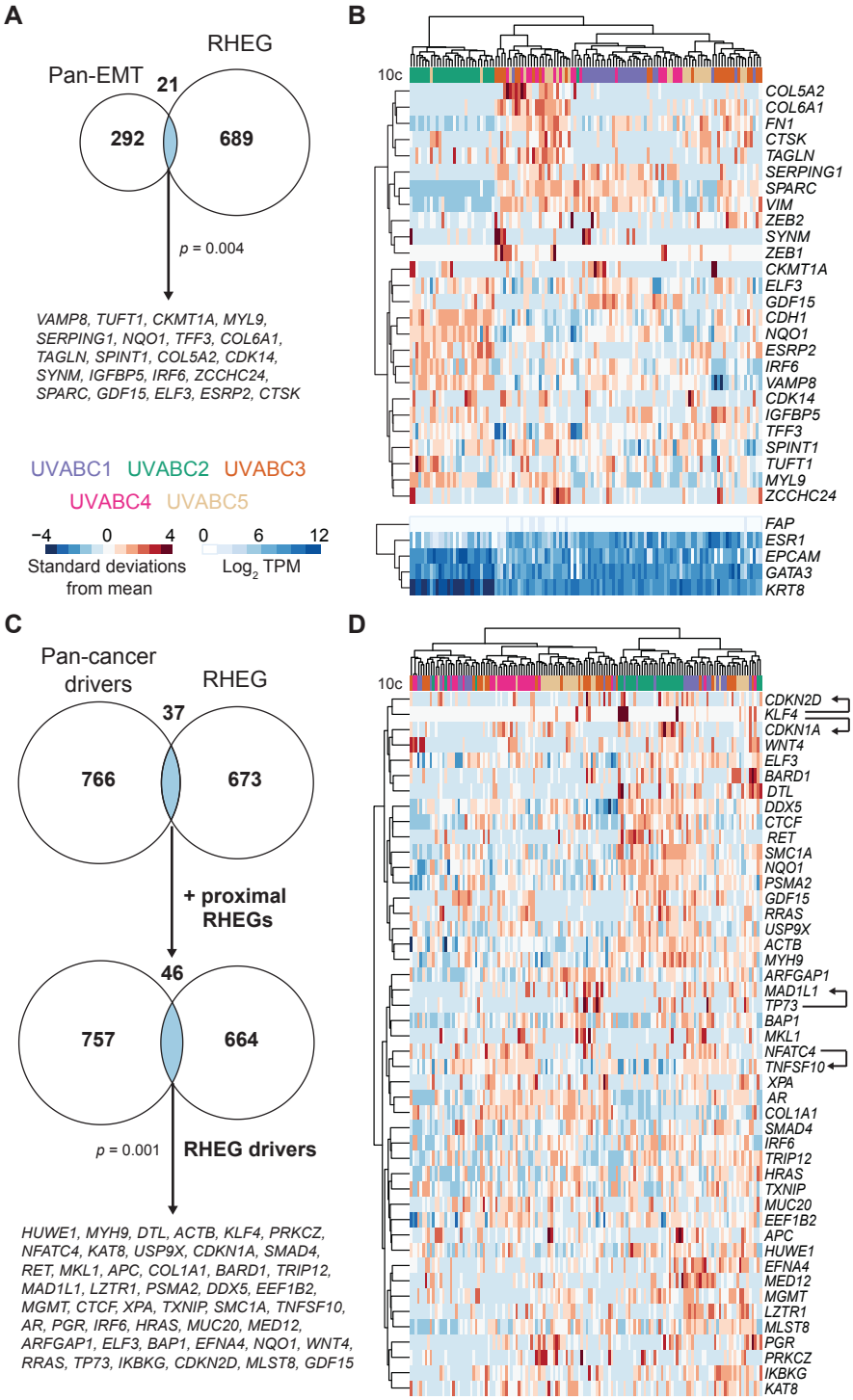
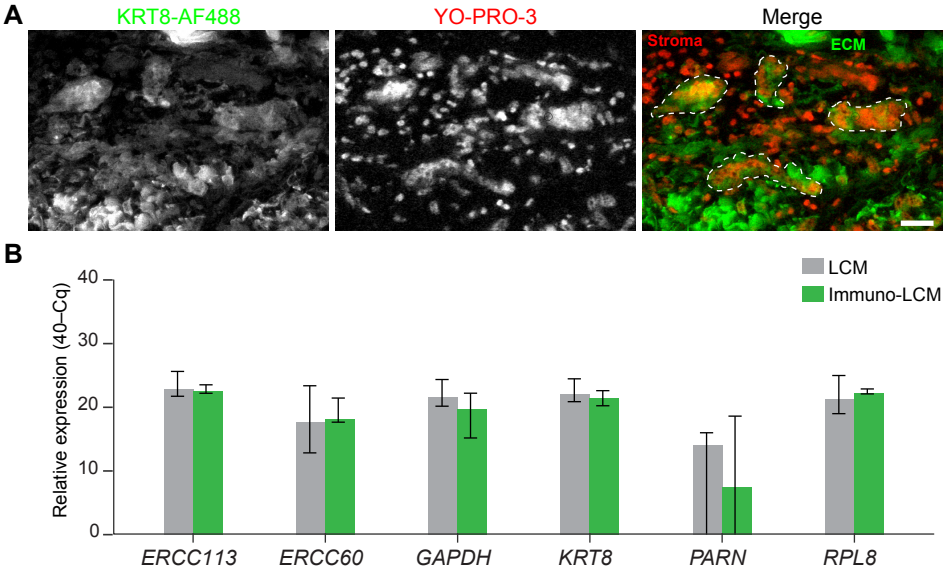

A
Pan-EMT
RHEG
21
292
689
p = 0.004
VAMP8, TUFT1, CKMT1A, MYL9, SERPING1, NQO1, TFF3, COL6A1, TAGLN, SPINT1, COL5A2, CDK14, SYNM, IGFBP5, IRF6, ZCCHC24, SPARC, GDF15, ELF3, ESRP2, CTSK
UVABC1 UVABC2 UVABC3
UVABC4 UVABC5
−4 0 4
Standard deviations from mean
0 6 12
Log2 TPM
B
10c
COL5A2
COL6A1
FN1
CTSK
TAGLN
SERPING1
SPARC
VIM
ZEB2
SYNM
ZEB1
CKMT1A
ELF3
GDF15
CDH1
NQO1
ESRP2
IRF6
VAMP8
CDK14
IGFBP5
TFF3
SPINT1
TUFT1
MYL9
ZCCHC24
FAP
ESR1
EPCAM
GATA3
KRT8
C
Pan-cancer drivers
RHEG
37
766
673
+ proximal RHEGs
46
757
664
p = 0.001
RHEG drivers
HUWE1, MYH9, DTL, ACTB, KLF4, PRKCZ, NFATC4, KAT8, USP9X, CDKN1A, SMAD4, RET, MKL1, APC, COL1A1, BARD1, TRIP12, MAD1L1, LZTR1, PSMA2, DDX5, EEF1B2, MGMT, CTCF, XPA, TXNIP, SMC1A, TNFSF10, AR, PGR, IRF6, HRAS, MUC20, MED12, ARFGAP1, ELF3, BAP1, EFNA4, NQO1, WNT4, RRAS, TP73, IKBKG, CDKN2D, MLST8, GDF15
D
10c
CDKN2D
KLF4
CDKN1A
WNT4
ELF3
BARD1
DTL
DDX5
CTCF
RET
SMC1A
NQO1
PSMA2
GDF15
RRAS
USP9X
ACTB
MYH9
ARFGAP1
MAD1L1
TP73
BAP1
MKL1
NFATC4
TNFSF10
XPA
AR
COL1A1
SMAD4
IRF6
TRIP12
HRAS
TXNIP
MUC20
EEF1B2
APC
HUWE1
EFNA4
MED12
MGMT
LZTR1
MLST8
PGR
PRKCZ
IKBKG
KAT8

Figure 6. RHEGs contain epithelial-to-mesenchymal transition (EMT) markers and driver genes for cancers other than breast. **A,** Venn diagram intersecting the UVABC RHEG set with a pan-cancer EMT signature (36). Shared genes are listed. **B,** Hierarchical clustering of the shared genes in **A** along with epithelial (*CDH1*) and mesenchymal (*VIM*, *FN1*, *ZEB1*, *ZEB2*) markers that were reliably detected by 10cRNA-seq. Stromal contamination was excluded by the relative abundance of the fibroblast marker *FAP* compared to the luminal markers *ESR1*, *EPCAM*, *GATA3*, and *KRT8*. **C,** Venn diagram intersecting the UVABC RHEG set with a pan-cancer set of driver genes (23,37). The intersection was updated to include proximal RHEGs as described in Supplementary Table ST2. Shared genes ("RHEG drivers") are listed. **D,** Hierarchical clustering of RHEG drivers. Arrows between co-clustering drivers indicate possibly direct transcription factor–target gene relationships as described in the text. For **A** and **C**, statistical significance of overlaps was assessed by the hypergeometric test.

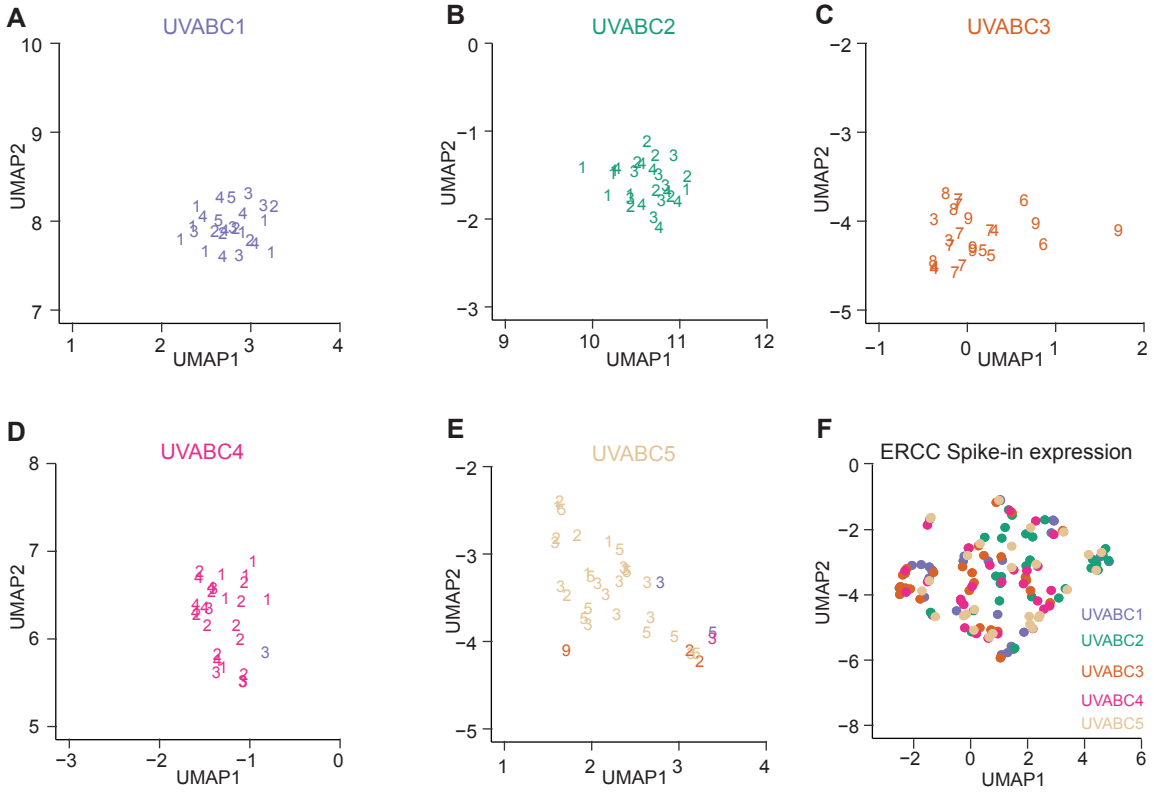
Table 1. Transcriptomic studies of intra-carcinoma cell heterogeneity from primary clinical cases.

Cancer type	Median stage [range]	Dissociation time	Carcinoma yield	Read depth per cell	Reference
Head and neck	T3N1 [T1N0–T4aN2c]	60 min	38%	1340K	(41)
Colon	T3N2 [T1n0–T4bN0]	30 min	28%	≥100K	(55)
Pancreas	IIb [Ib–IV]	45 min	7.8%	50K	(77)
	T2N1 [T1cN0–T4N2]	40 min	6.3%	50K	(78)
Breast	T2N0 [T1N0–T3N1]	n.a.	62%	5800K	(28)
	T2N0 [T0N0–T2N3]	60 min	62%	50K	(42)
Luminal breast	T1N0 [T1N0-T2N1]	0 min	~100%	604K	This study

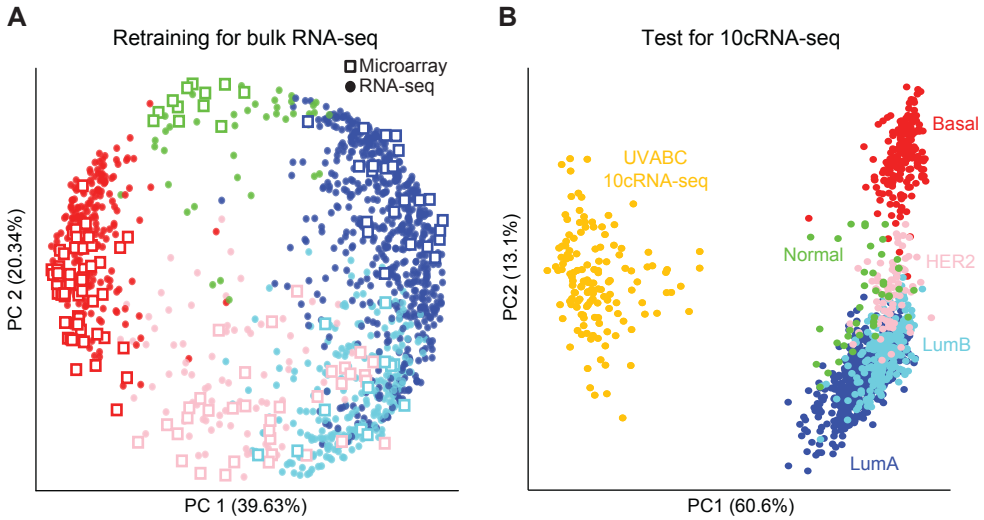
n.a., not available.



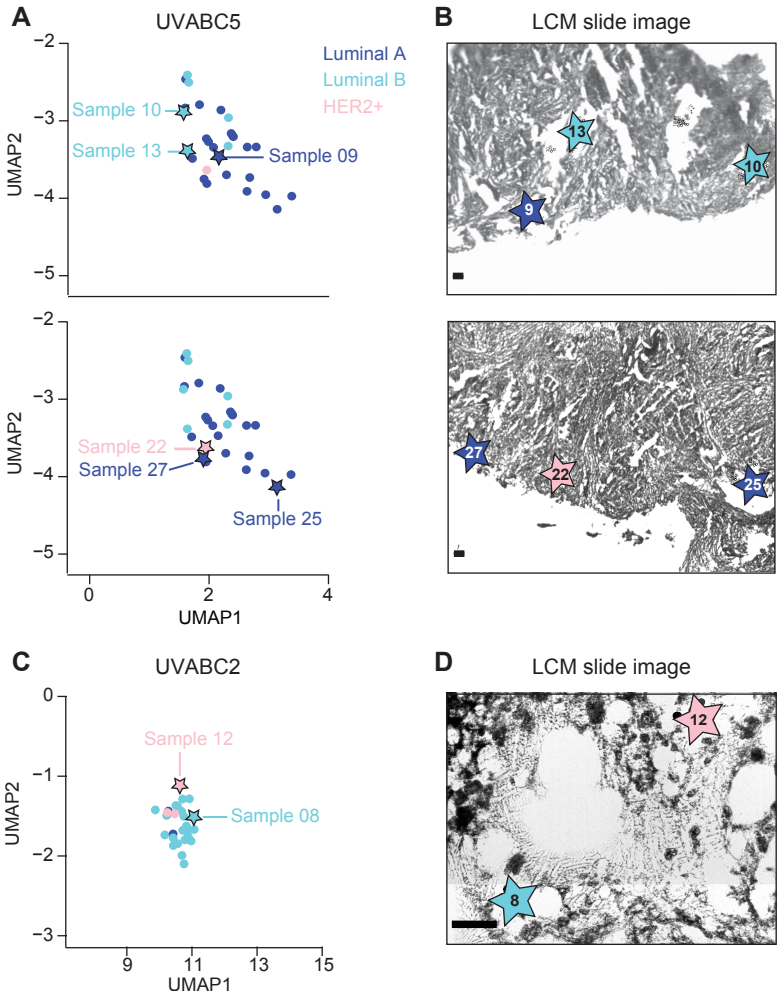
Supplementary Figure S1. Immuno-LCM capture of 10-cell samples does not affect gene detection. **A,** Rapid immunostaining of UVABC3 with an Alexa Fluor 488 (AF488)-conjugated antibody recognizing KRT8 (left) combined with YO-PRO-3 to stain all nuclei (middle). Epithelial cells were identified by dual red-green staining (right, dashed lines) compared to stromal cells that are KRT8-negative and autofluorescent extracellular matrix that is free of nuclear staining. Scale bar is 25 µm. **B,** Relative abundance for the indicated transcripts as measured by quantitative PCR in UVABC4. Two exogenously spiked-in RNA transcripts (*ERCC113*, *ERCC60*) were quantified along with four endogenous genes. Data are shown as the median inverse quantification cycle (40–Cq) ± range from n = 4 replicates.



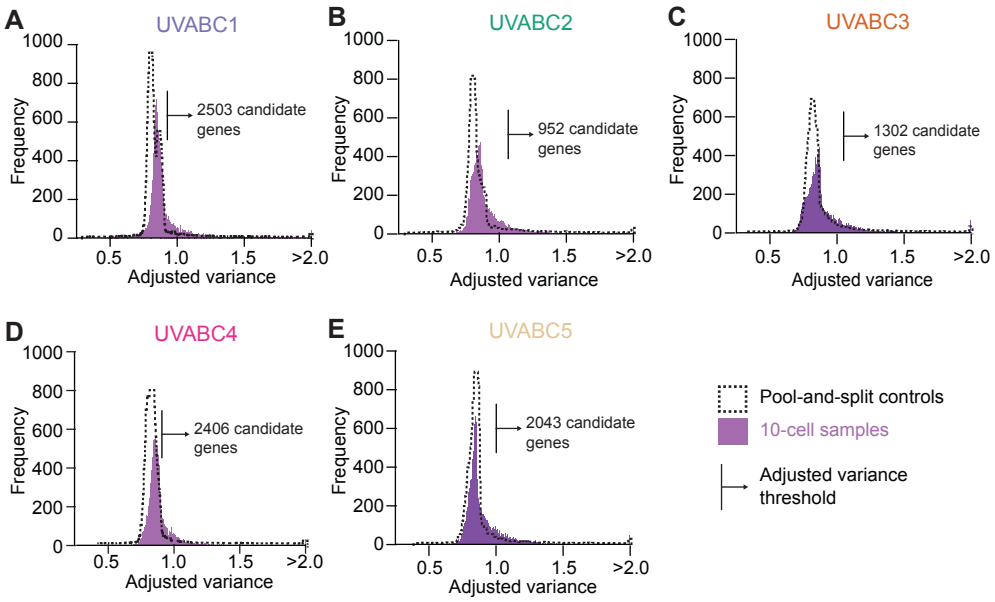
Supplementary Figure S2. Clustering of 10cRNA-seq data by tumor does not arise from batch effects. **A–E,** Enlarged UMAP embedding for UVABC tumors from **Fig. 2A**, with different batches of sample collections annotated by number. **F,** UMAP embedding of the UVABC cohort based only on the ERCC spike-in transcripts of 10cRNA-seq samples.



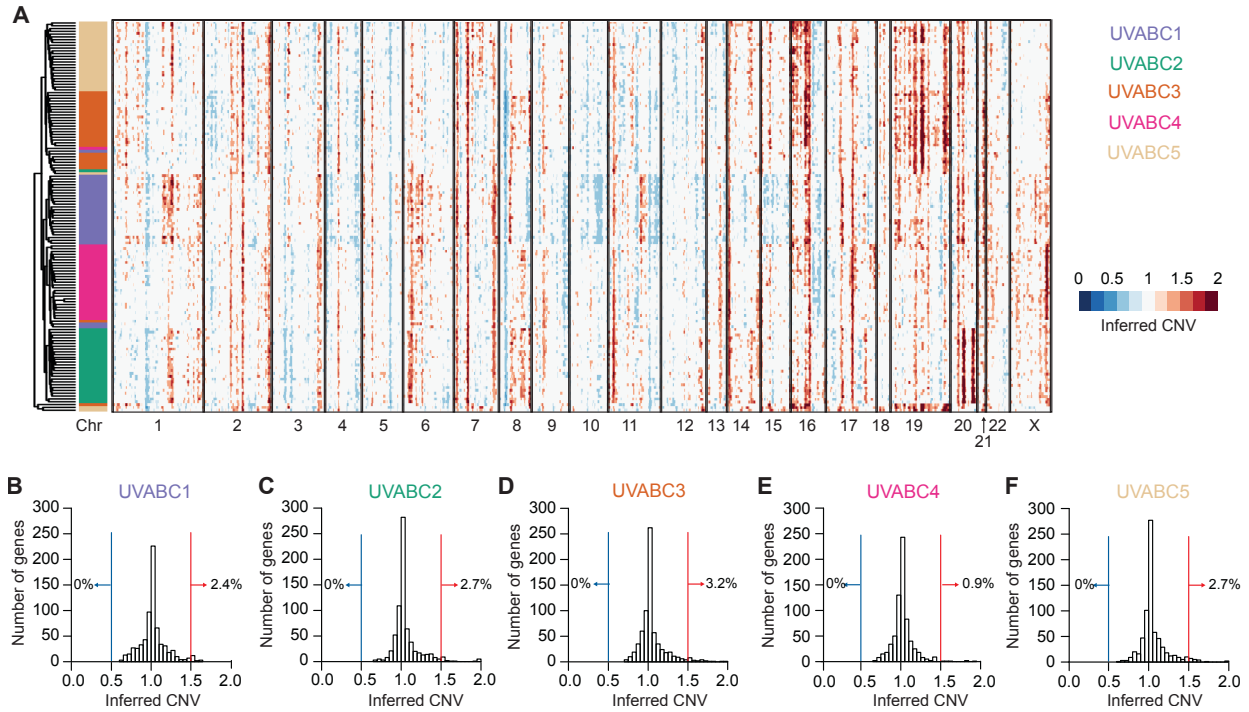
Supplementary Figure S3. Microarray-based PAM50 classification is not adaptable to 10cRNA-seq. **A,** Principal component plot adapting the microarray-based PAM50 predictor (27) to bulk RNA-seq data of breast tumors from The Cancer Genome Atlas (TCGA) (26) (see Materials and Methods). Similar projections of the different data types indicate successful data fusion: 77% of RNA-seq samples were correctly subtyped, with an average confidence score of 0.99. **B,** Principal component plot of UVABC 10cRNA-seq observations together with TCGA tumors subtyped with the adapted PAM50 classification of **A**. 10cRNA-seq projections are distinct from all classified subtypes of TCGA tumors, even though dispersion along PC2 indicates underlying subtype differences.



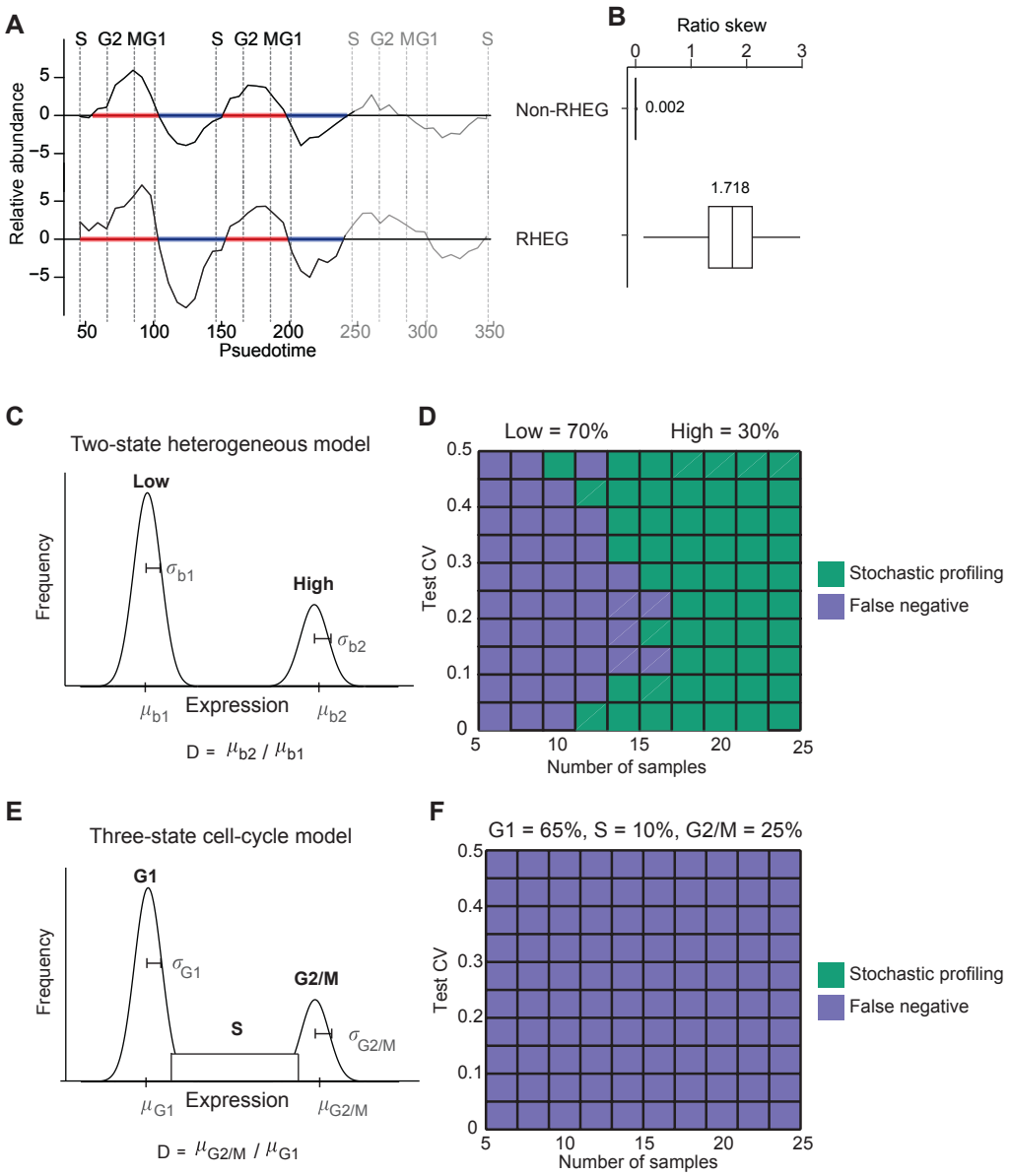
Supplementary Figure S4. Different molecular subtypes are assigned to tumor cells microdissected from the same cryosection. **A,** Enlarged UMAP embedding for UVABC5 from **Fig. 2C**, highlighting six 10cRNA-seq samples (stars) obtained together in two separate cryosections. **B,** Low-magnification grayscale LCM slide images indicating the regions microdissected in the UVABC5 cryosections. **C,** Enlarged UMAP embedding for UVABC2 from **Fig. 2C**, highlighting two 10cRNA-seq samples (stars) obtained together in one cryosection. **D,** High-magnification grayscale LCM slide image indicating the regions microdissected in the UVABC2 cryosection. Scale bar is 50 µm (**B** and **D**).



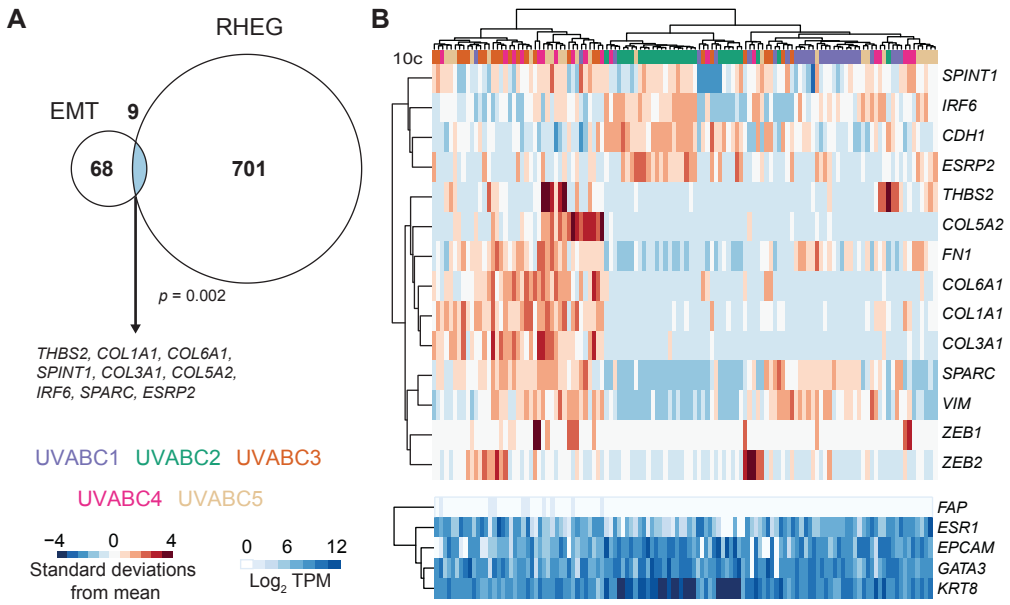
Supplementary Figure S5. Abundance-dependent overdispersion statistics of individual cases in the UVABC cohort. **A–E,** Distribution of adjusted variance scores for each gene measured transcriptomically in separate 10-cell samples (purple) compared to pool-and-split controls estimating technical variation (black dashed). The arrow indicates the upper 5th percentile of adjusted variance for the pool-and-split controls, which is used as the cutoff for 10-cell samples.



Supplementary Figure S6. Most candidate heterogeneities do not reside in loci with inferred copy-number variations (CNVs). **A,** Chromosomal gains and losses predicted from 10cRNA-seq data by inferCNV (30). RNA-seq data from normal human luminal breast tissue obtained through GTEx (29) was used as the reference transcriptome. Gains in 1q (UVABC1, UVABC2) and 8q (UVABC2, UVABC4) and losses in 8p (all) and 16q (UVABC1, UVABC4) are characteristic of luminal A breast tumors (79). **B–F,** Distribution of inferred CNVs corresponding to the candidate heterogeneities of individual cases in the UVABC cohort. Percentage of transcripts with inferCNV scores suggesting gain (red) or loss (blue) is shown.


A
S G2 MG1 S G2 MG1 S G2 MG1 S
Relative abundance
Psuedotime
B
Ratio skew
Non-RHEG
0.002
RHEG
1.718
C
Two-state heterogeneous model
Low
High
Frequency
Expression
D = μb2 / μb1
D
Low = 70%
High = 30%
Test CV
Number of samples
Stochastic profiling
False negative
E
Three-state cell-cycle model
G1
S
G2/M
Frequency
Expression
D = μG2/M / μG1
F
G1 = 65%, S = 10%, G2/M = 25%
Test CV
Number of samples
Stochastic profiling
False negative

Supplementary Figure S7. Periodically cycling transcripts are disfavored by stochastic profiling. **A,** Asymmetric cell-cycle oscillations in the RHEG *DCTPP1* compared to the non-RHEG *HJURP*. Microarray data from synchronized HeLa cells (33) and pseudotime estimates were obtained from Cyclebase 3.0 (32). Time intervals above (red) and time (blue) the midpoint of each transcript are shown. **B,** Asymmetry of cycling RHEGs quantified by ratio skew and compared to non-RHEG cycling transcripts (see Materials and Methods). Data are shown as boxplots from n = 1000 bootstrapping runs. **C,** Abstraction of a two-state regulatory heterogeneity. Parameters of the probability distribution are described elsewhere (19,20,22). **D,** Monte-Carlo simulations (19) of stochastic profiling in the two-state case. False-negative regimes are marked when a two-state heterogeneity is not detected in the 10-cell pool. **E,** Abstraction of a three-state cell-cycle model. A uniform S-phase interval is added in between the first and second regulatory states modeling G1 and G2/M phases, and the fractional proportions are updated accordingly (see Materials and Methods). **F,** Monte-Carlo simulation of stochastic profiling in the three-state case. The false-negative regime marks three-state heterogeneities that are not detected in the 10-cell pool. For **D** and **F**, the following simulation parameters were used: $D = 3$, $\sigma_{b1} = \sigma_{b2} = \sigma_{G1} = \sigma_{G2} = 0.2$. Code for the three-state cell-cycle model is available in Supplementary File S2.



Supplementary Figure S8. RHEGs are enriched for additional epithelial-to-mesenchymal transition (EMT) signatures. **A,** Venn diagram intersecting the UVABC RHEG set with an independently derived EMT signature (24) from that of **Fig. 6A**. Shared genes are listed. **B,** Hierarchical clustering of the shared genes in A along with epithelial (*CDH1*) and mesenchymal (*VIM*, *FN1*, *ZEB1*, *ZEB2*) markers that were reliably detected by 10cRNA-seq. Stromal contamination was excluded by the relative abundance of the fibroblast marker *FAP* compared to the luminal markers *ESR1*, *EPCAM*, *GATA3*, and *KRT8*.

Supplementary Table ST1. Early stage luminal tumors profiled in this study.

Sample	Age	Race	Stage	Grade	ER	PR	HER2	Mitoses (per hpf)	Tumor infiltrating lymphocytes foci (per hpf)	Desmoplastic reaction
UVABC1	76	Caucasian	1	3	+	+	—	21/10	2/5	Marked
UVABC2	52	Caucasian	2	3	+	+	—	32/10	2/5	Moderate
UVABC3	63	Caucasian	1	3	+	+	—	11/10	5/5	Marked
UVABC4	63	Caucasian	1	2	+	+	—	2/10	1/5	Marked
UVABC5	59	Caucasian	1	1	+	+	—	2/10	3/5	Marked

hpf = high powered field (400x magnification)

Supplementary Table ST2. Multiple RHEGs are proximal to established cancer driver genes.

RHEG	Driver gene	Cancer type(s)	Proximal relationship
EFNA4	*EPHA4*	LUAD	EPHA4 signals through EFNA4
GDF15	*RET*	PANCAN, THCA, SKCM	RET is a coreceptor for GDF15
NQO1	*NFE2L2*	PANCAN, LUSC, LIHC	NFE2L2 activity is marked by *NQO1* abundance
WNT4	*WNT5A*	PRAD	WNT5A and WNT4 are both non-canonical Wnt ligands
RRAS	*KRAS*	PANCAN, COAD, LUAD, PAAD, UCEC, ESCA	KRAS and RRAS are both in the Ras family
TP73	*TP53*	PANCAN, BLCA, BRCA, GBM, COAD, ESCA, HNSC, KIRC, LIHC, LUAD, LUSC, DLBC, OV, PAAD, PRAD, SKCM, STAD, UCEC, SARC	TP53 and TP73 are in the same family of transcription factors
IKBKG	*IKBKB*	DLBC	IKBKB and IKBKG are in the same IKK complex
CDKN2D	*CDKN2A*	PANCAN, HNSC, ESCA, LUSC, PAAD, SKCM	CDKN2A and CDKN2D are related CDK inhibitors
MLST8	*MTOR*	KIRC	MTOR and MLST8 are in the same MTORC complex

Acknowledgments

We thank Emily Farber and Suna Onengut-Gumuscu at the UVA Center for Public Health Genomics for RNA-seq library preparation and sequencing, Craig Rumpel and Angela Miller at the UVA Biorepository and Tissue Research Facility for LCM maintenance and histology services, UVA Research Computing for high-performance computing access and consulting, Henry Pritchard for assistance with NCBI GEO deposition, and Dylan Schaff, Francine Garrett-Bakelman, and Cheryl Borgman for critical evaluation of the manuscript. This work was supported by the National Institutes of Health #R01-CA194470 (K.A.J.), the David & Lucile Packard Foundation #2009-34710 (K.A.J.), a Wagner Fellowship (S.S.), the UVA Medical Scientist Training Program (S.S.), and a UVA Cancer Center support grant #P30-CA044579.

References

1. Surveillance, Epidemiology, and End Results (SEER) Program. <https://seer.cancer.gov>.
2. Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science **2015**;347:78-81
3. Medina D. The mammary gland: a unique organ for the study of development and tumorigenesis. J Mammary Gland Biol Neoplasia **1996**;1:5-19
4. Nguyen-Ngoc KV, Cheung KJ, Brenot A, Shamir ER, Gray RS, Hines WC*, et al.* ECM microenvironment regulates collective migration and local dissemination in normal and malignant mammary epithelium. Proc Natl Acad Sci U S A **2012**;109:E2595-604
5. Chaudhuri O, Koshy ST, Branco da Cunha C, Shin JW, Verbeke CS, Allison KH*, et al.* Extracellular matrix stiffness and composition jointly regulate the induction of malignant phenotypes in mammary epithelium. Nat Mater **2014**;13:970-8
6. Cerchiari AE, Garbe JC, Jee NY, Todhunter ME, Broaders KE, Peehl DM*, et al.* A strategy for tissue self-organization that is robust to cellular heterogeneity and plasticity. Proc Natl Acad Sci U S A **2015**;112:2287-92
7. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J*, et al.* Tumour evolution inferred by single-cell sequencing. Nature **2011**;472:90-4
8. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D*, et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med **2012**;366:883-92
9. Nieto MA, Huang RY, Jackson RA, Thiery JP. Emt: 2016. Cell **2016**;166:21-45
10. Davis A, Gao R, Navin N. Tumor evolution: Linear, branching, neutral or punctuated? Biochim Biophys Acta Rev Cancer **2017**;1867:151-61
11. Janes KA. Single-cell states versus single-cell atlases - two classes of heterogeneity that differ in meaning and method. Curr Opin Biotechnol **2016**;39:120-5
12. Suva ML, Tirosh I. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. Mol Cell **2019**;75:7-12
13. Svensson V, Natarajan KN, Ly LH, Miragaia RJ, Labalette C, Macaulay IC*, et al.* Power analysis of single-cell RNA-sequencing experiments. Nat Methods **2017**;14:381-7
14. van den Brink SC, Sage F, Vertesy A, Spanjaard B, Peterson-Maduro J, Baron CS*, et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat Methods **2017**;14:935-6
15. Machado L, Esteves de Lima J, Fabre O, Proux C, Legendre R, Szegedi A*, et al.* In Situ Fixation Redefines Quiescence and Early Activation of Skeletal Muscle Stem Cells. Cell Rep **2017**;21:1982-93
16. Singh S, Wang L, Schaff DL, Sutcliffe MD, Koeppel AF, Kim J*, et al.* In situ 10-cell RNA sequencing in tissue and tumor biopsy samples. Sci Rep **2019**;9:4836
17. Nichterwitz S, Chen G, Aguila Benitez J, Yilmaz M, Storvall H, Cao M*, et al.* Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. Nat Commun **2016**;7:12139
18. Chen J, Suo S, Tam PP, Han JJ, Peng G, Jing N. Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. Nat Protoc **2017**;12:566-80
19. Wang L, Janes KA. Stochastic profiling of transcriptional regulatory heterogeneities in tissues, tumors and cultured cells. Nat Protoc **2013**;8:282-301
20. Janes KA, Wang CC, Holmberg KJ, Cabral K, Brugge JS. Identifying single-cell molecular programs by stochastic profiling. Nat Methods **2010**;7:311-7

21. Wang L, Brugge JS, Janes KA. Intersection of FOXO- and RUNX1-mediated gene expression programs in single breast epithelial cells during morphogenesis and tumor progression. Proc Natl Acad Sci U S A **2011**;108:E803-12
22. Bajikar SS, Fuchs C, Roller A, Theis FJ, Janes KA. Parameterizing cell-to-cell regulatory heterogeneities via stochastic transcriptional profiles. Proc Natl Acad Sci U S A **2014**;111:E626-35
23. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A*, et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell **2018**;173:371-85 e18
24. Mak MP, Tong P, Diao L, Cardnell RJ, Gibbons DL, William WN*, et al.* A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition. Clin Cancer Res **2016**;22:609-20
25. Consortium ITP-CAoWG. Pan-cancer analysis of whole genomes. Nature **2020**;578:82-93
26. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A*, et al.* Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell **2015**;163:506-19
27. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T*, et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol **2009**;27:1160-7
28. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT*, et al.* Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun **2017**;8:15081
29. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science **2015**;348:648-60
30. inferCNV of the Trinity CTAT Project. <https://github.com/broadinstitute/inferCNV>.
31. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H*, et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science **2014**;344:1396-401
32. Santos A, Wernersson R, Jensen LJ. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. Nucleic Acids Res **2015**;43:D1140-4
33. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE*, et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell **2002**;13:1977-2000
34. Aitken SJ, Ibarra-Soria X, Kentepozidou E, Flicek P, Feig C, Marioni JC*, et al.* CTCF maintains regulatory homeostasis of cancer pathways. Genome Biol **2018**;19:106
35. Takaku M, Grimm SA, Roberts JD, Chrysovergis K, Bennett BD, Myers P*, et al.* GATA3 zinc finger 2 mutations reprogram the breast cancer transcriptional network. Nat Commun **2018**;9:1059
36. Tan TZ, Miow QH, Miki Y, Noda T, Mori S, Huang RY*, et al.* Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. EMBO Mol Med **2014**;6:1279-93
37. Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G*, et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature **2020**;578:102-11
38. Matsuno RK, Anderson WF, Yamamoto S, Tsukuma H, Pfeiffer RM, Kobayashi K*, et al.* Early- and late-onset breast cancer types among women in the United States and Japan. Cancer Epidemiol Biomarkers Prev **2007**;16:1437-42
39. Fend F, Emmert-Buck MR, Chuaqui R, Cole K, Lee J, Liotta LA*, et al.* Immuno-LCM: laser capture microdissection of immunostained frozen sections for mRNA analysis. Am J Pathol **1999**;154:61-6

40. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG*, et al.* Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol **2018**
41. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S*, et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell **2017**;171:1611-24 e24
42. Yin J, Li Z, Yan C, Fang E, Wang T, Zhou H*, et al.* Comprehensive analysis of immune evasion in breast cancer by single-cell RNA-seq. bioRxiv **2018**:368605
43. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA*, et al.* Molecular portraits of human breast tumours. Nature **2000**;406:747-52
44. Gendoo DM, Ratanasirigulchai N, Schroeder MS, Pare L, Parker JS, Prat A*, et al.* genefu: Computation of Gene Expression-Based Signatures in Breast Cancer. **2020**:R package version 2.18.1
45. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM*, et al.* From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res **2014**;24:496-510
46. Picelli S, Bjorklund AK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods **2013**;10:1096-8
47. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods **2014**;11:740-2
48. Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL*, et al.* Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat Methods **2016**;13:241-4
49. Schaff DL, Singh S, Kim KB, Sutcliffe MD, Park KS, Janes KA. Fragmentation of small-cell lung cancer regulatory states in heterotypic microenvironments. Cancer Res **2020**:co-submitted
50. Sutcliffe MD, Galvao RP, Wang L, Kim J, Singh S, Zong H*, et al.* Single-cell bottlenecks and dead-ends during glioma premalignancy. Cancer Res **2020**:co-submitted
51. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ*, et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol **2015**;33:155-60
52. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M*, et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell **2015**;161:1202-14
53. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW*, et al.* The life history of 21 breast cancers. Cell **2012**;149:994-1007
54. Shlien A, Raine K, Fuligni F, Arnold R, Nik-Zainal S, Dronov S*, et al.* Direct Transcriptional Consequences of Somatic Mutation in Breast Cancer. Cell Rep **2016**;16:2032-46
55. Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL*, et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet **2017**;49:708-18
56. Hu Z, Artibani M, Alsaadi A, Wietek N, Morotti M, Shi T*, et al.* The Repertoire of Serous Ovarian Cancer Non-genetic Heterogeneity Revealed by Single-Cell Sequencing of Normal Fallopian Tube Epithelial Cells. Cancer Cell **2020**;37:226-42 e7
57. Savci-Heijink CD, Halfwerk H, Hooijer GKJ, Koster J, Horlings HM, Meijer SL*, et al.* Epithelial-to-mesenchymal transition status of primary breast carcinomas and its correlation with metastatic behavior. Breast Cancer Res Treat **2019**;174:649-59
58. Wang Q, Zhou Y, Weiss HL, Chow CW, Evers BM. NFATc1 regulation of TRAIL expression in human intestinal cells. PLoS One **2011**;6:e19882

59. Zhang W, Geiman DE, Shields JM, Dang DT, Mahatan CS, Kaestner KH*, et al.* The gut-enriched Kruppel-like factor (Kruppel-like factor 4) mediates the transactivating effect of p53 on the p21WAF1/Cip1 promoter. J Biol Chem **2000**;275:18391-8
60. Aksoy I, Giudice V, Delahaye E, Wianny F, Aubry M, Mure M*, et al.* Klf4 and Klf5 differentially inhibit mesoderm and endoderm differentiation in embryonic stem cells. Nat Commun **2014**;5:3719
61. Santos Guasch GL, Beeler JS, Marshall CB, Shaver TM, Sheng Q, Johnson KN*, et al.* p73 Is Required for Ovarian Follicle Development and Regulates a Gene Network Involved in Cell-to-Cell Adhesion. iScience **2018**;8:236-49
62. Fischer M. Census and evaluation of p53 target genes. Oncogene **2017**;36:3943-56
63. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S*, et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell **2014**;158:929-44
64. Wagner J, Rapsomaniki MA, Chevrier S, Anzeneder T, Langwieder C, Dykgers A*, et al.* A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. Cell **2019**;177:1330-45 e18
65. Gut G, Herrmann MD, Pelkmans L. Multiplexed protein maps link subcellular organization to cellular states. Science **2018**;361
66. Stahl PL, Salmen F, Vickovic S, Lundmark A, Navarro JF, Magnusson J*, et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science **2016**;353:78-82
67. Wei SC, Fattet L, Tsai JH, Guo Y, Pai VH, Majeski HE*, et al.* Matrix stiffness drives epithelial-mesenchymal transition and tumour metastasis through a TWIST1-G3BP2 mechanotransduction pathway. Nat Cell Biol **2015**;17:678-88
68. Kumar MP, Du J, Lagoudas G, Jiao Y, Sawyer A, Drummond DC*, et al.* Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. Cell Rep **2018**;25:1458-68 e4
69. Li PX, Wong J, Ayed A, Ngo D, Brade AM, Arrowsmith C*, et al.* Placental transforming growth factor-beta is a downstream mediator of the growth arrest and apoptotic response of tumor cells to DNA damage and p53 overexpression. J Biol Chem **2000**;275:20127-35
70. Mullican SE, Lin-Schmidt X, Chin CN, Chavez JA, Furman JL, Armstrong AA*, et al.* GFRAL is the receptor for GDF15 and the ligand promotes weight loss in mice and nonhuman primates. Nat Med **2017**;23:1150-7
71. Chang GW, Hsiao CC, Peng YM, Vieira Braga FA, Kragten NA, Remmerswaal EB*, et al.* The Adhesion G Protein-Coupled Receptor GPR56/ADGRG1 Is an Inhibitory Receptor on Human NK Cells. Cell Rep **2016**;15:1757-70
72. Lacher MD, Bauer G, Fury B, Graeve S, Fledderman EL, Petrie TD*, et al.* SV-BR-1-GM, a Clinically Effective GM-CSF-Secreting Breast Cancer Cell Line, Expresses an Immune Signature and Directly Activates CD4(+) T Lymphocytes. Front Immunol **2018**;9:776
73. Yao HP, Zhou YQ, Zhang R, Wang MH. MSP-RON signalling in cancer: pathogenesis and therapeutic potential. Nat Rev Cancer **2013**;13:466-81
74. Yang Y, Hsu JM, Sun L, Chan LC, Li CW, Hsu JL*, et al.* Palmitoylation stabilizes PD-L1 to promote breast tumor growth. Cell Res **2019**;29:83-6
75. Tsujikawa T, Kumar S, Borkar RN, Azimi V, Thibault G, Chang YH*, et al.* Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis. Cell Rep **2017**;19:203-17
76. van Seijen M, Lips EH, Thompson AM, Nik-Zainal S, Futreal A, Hwang ES*, et al.* Ductal carcinoma in situ: to treat or not to treat, that is the question. Br J Cancer **2019**;121:285-92

77. Elyada E, Bolisetty M, Laise P, Flynn WF, Courtois ET, Burkhart RA*, et al.* Cross-Species Single-Cell Analysis of Pancreatic Ductal Adenocarcinoma Reveals Antigen-Presenting Cancer-Associated Fibroblasts. Cancer Discov **2019**;9:1102-23
78. Peng J, Sun BF, Chen CY, Zhou JY, Chen YS, Chen H*, et al.* Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res **2019**;29:725-38
79. Ciriello G, Sinha R, Hoadley KA, Jacobsen AS, Reva B, Perou CM*, et al.* The molecular diversity of Luminal A breast tumors. Breast Cancer Res Treat **2013**;141:409-20