Single-Cell Expression Analysis of Ductal Carcinoma *In Situ* Identifies Complex Genotypic–Phenotypic Relationships Altering Epithelial Composition

Xiaodi Qin[1], Siri H. Strand[2], Marissa R. Lee[1], Aashrith Saraswathibhatla[3], David G.P. van IJzendoorn[2], ChunFang Zhu[2], Sujay Vennam[2], Sushama Varma[2], Allison Hall[4], Rachel E. Factor[4], Lorraine King[5], Lunden Simpson[5], Xiaoke Luo[6], Graham A. Colditz[6], Shu Jiang[6], Ovijit Chaudhuri[3], E. Shelley Hwang[5], Jeffrey R. Marks[5], Kouros Owzar[1,7], and Robert B. West[2]

ABSTRACT

Ductal carcinoma *in situ* (DCIS) is a risk factor for subsequent invasive breast cancer (IBC). To identify events in DCIS that lead to invasive cancer, we performed single-cell RNA sequencing on DCIS lesions and matched normal breast tissue. Inferred copy-number variation was used to identify neoplastic epithelial cells from clinical specimens, which contained a mixture of DCIS and normal ducts. Phylogenetic analysis demonstrated intratumoral clonal heterogeneity that was associated with significant gene expression differences. Classification of epithelial cells into mammary cell states revealed that subclones contained a mixture of cell states, suggesting an ongoing pattern of differentiation after neoplastic transformation. Cell state proportions were significantly different based on estrogen receptor expression, with estrogen receptor–negative DCIS more closely resembling the distribution in the normal breast, particularly with respect to cells with basal characteristics. Specific alterations in cell state proportions were associated with progression to invasive cancer in a cohort of DCIS with longitudinal outcome. Ongoing transcription of key basement membrane (BM) genes occurred in specific subsets of epithelial cell states, including basal/myoepithelial, which are diminished in DCIS. In the transition to IBC, the BM protein laminin, but not COL4, was altered in DCIS adjacent to invasion. Loss of COL4, but not laminin, in an *in vitro* DCIS model led to an invasive phenotype. These findings suggest that the process of invasion is a loss-of-function event due to an imbalance in critical cell populations essential for BM integrity rather than a gain of an invasive phenotype by neoplastic cells.

Significance: Single-cell analyses reveal ductal carcinoma *in situ* comprises multiple genetic clones with significant phenotypic diversity and link alterations in epithelial cell states and basement membrane integrity with invasive breast cancer progression.

Introduction

Ductal carcinoma *in situ* (DCIS) of the breast is a noninvasive condition commonly identified through mammographic screening. A primary diagnosis of DCIS carries little mortality risk on its own, but its presence is a risk factor for subsequent clonally related invasive breast cancer (IBC; refs. 1–5). DCIS is a neoplastic proliferation of mammary epithelial cells within an intact basement membrane (BM). These lesions share virtually all of the genomic alterations found in IBC, and therefore no specific genetic events have been associated with disease progression (6), although there are expression signatures that predict progression and benefit of radiotherapy (7–9). Mechanisms of invasion remain poorly understood. Genomic analysis at the single epithelial cell level indicates that invasion is commonly polyclonal (10). This is consistent with no single genomic event being responsible for the acquisition of the invasive phenotype. Processes that have been proposed as mechanisms of invasion include active degradation of the BM by neoplastic cells, loss of myoepithelial cells, which may serve as a cellular barrier to invasion, and mechanical pressures from the proliferative and expanding duct (11–15).

Single-cell RNA sequencing (scRNA-seq) has emerged as a valuable tool to understand tissue composition and expression at single-cell and whole-transcriptome resolutions. A recent application of this approach to DCIS concluded that significant intratumoral and intertumoral heterogeneity exists (16); however, insights into the genesis of DCIS heterogeneity was limited by the lack of patient-matched normal samples. As part of the Human

[1]Duke Cancer Institute, Duke University School of Medicine, Durham, North Carolina. [2]Department of Pathology, Stanford University School of Medicine, Stanford, California. [3]Department of Mechanical Engineering, Stanford University, Stanford, California. [4]Department of Pathology, Duke University School of Medicine, Durham, North Carolina. [5]Department of Surgery, Duke University School of Medicine, Durham, North Carolina. [6]Department of Surgery, Washington University School of Medicine, St. Louis, Missouri. [7]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina.

X. Qin and S.H. Strand contributed equally as co-first authors to this article.

J.R. Marks, K. Owzar, and R.B. West contributed equally to this article.

Corresponding Authors: Robert B. West, Department of Pathology, Stanford University Medical Center, Room L215A, 300 Pasteur Drive, Stanford, CA 94305. E-mail: rbwest@stanford.edu; Jeffrey R. Marks, Division of Surgical Sciences, Department of Surgery, Duke University School of Medicine, LSRC Room B216, 450 Research Drive, Durham, NC 27705. E-mail: jeffrey.marks@duke.edu; and Kouros Owzar, Division of Integrative Genomics, Duke Center for Statistical Genetics and Genomic, Duke University School of Medicine, 2424 Erwin Road, Suite 1102, Office 7078, Durham, NC 27705. E-mail: kouros.owzar@duke.edu

Cancer Res 2025;85:2302–19

doi: 10.1158/0008-5472.CAN-24-3023

Tumor Atlas Network (HTAN), we focused on changes in cellular composition and phenotypes that occur between normal breast and DCIS starting from scRNA-seq data generated on DCIS lesions and matched synchronous normal breast tissue. We observed transcriptome heterogeneity, which accompanied intratumoral clonality, and significant alterations in epithelial cell states, including loss of specific epithelial cells that synthesize the BM. We extended the study to bulk RNA-seq data sets and analyses of DCIS and risk of progression to IBC in both patient data and an *in vitro* model. We find that BM loss can occur in DCIS adjacent to invasion and that decreasing BM components can lead to the invasive phenotype. We hypothesize that the altered cell composition of DCIS ducts leads to a loss of BM integrity and promotes subsequent invasion.

Materials and Methods

Institutional approval was obtained for the study (Stanford IRB-48262 and Duke IRB-Pro00100739). For the prospective sample set 1, written consent was obtained for all patients.

Sample sets

Sample set 1 (scRNA-seq set) consisted of DCIS ($n = 16$), IBC ($n = 2$), and synchronous normal tissues ($n = 12$) from 16 patients undergoing mastectomy at Duke and Stanford centers, collected between 2019 and 2023. Due to low sample size, the invasive samples were not included in these analyses. The demographic characteristics of patients analyzed ($n = 14$) are provided in Supplementary Table S1. Both the DCIS and normal specimens were bisected, with one half used for scRNA-seq, and the other half was formalin-fixed and paraffin-embedded (FFPE). Sections from the FFPE samples were used for histologic examination to confirm the presence and extent of DCIS and normal breast epithelium. The list of samples generated is provided in Supplementary Table S1.

Sample set 2 (matched and synchronous breast specimens) consisted of FFPE tissue samples from 42 patients operated on at Stanford Hospital with matched and synchronous areas of normal breast, DCIS, and IBC. The list of samples generated is provided in Supplementary Table S1.

Sample set 3 (DCIS with longitudinal outcome) consisted of 232 patients from the combined TBCRC and RAHBT cohorts included in the DCIS HTAN (7). These samples are the subset derived from patients with either no recurrence ($n = 163$) or IBC progression ($n = 69$). The list of published sample identifiers that were used in this study is provided in Supplementary Table S1.

Sample set 4 (DCIS with microinvasion) consisted of FFPE specimens with DCIS that exhibited areas of microinvasive cancer (<1 mm) collected from 13 patients at Stanford Hospital.

scRNA-seq assay

Fresh tissue samples were collected within 1 hour after devascularization and were immediately minced on ice before suspension in MACS Tissue Storage Solution (Miltenyi, cat. #130-100-008) in a 2-mL cryovial (Sarstedt, cat. #72.694.406) and stored at −80°C while awaiting pathologic confirmation of DCIS in the matching DCIS FFPE hematoxylin and eosin tissue section. Following confirmation, minced tissue samples were thawed in water bath and transferred to a gentleMACS C tube (Miltenyi, cat. #130-093-237) with enzymes from Miltenyi Human Tumor Dissociation Kit (130-095-929) according to protocol using gentleMACS Octo Dissociator with Heaters (130-096-427) program *37C_h_TDK_3*. Cells were washed with 3 mL Gibco RPMI 1640 Medium (Thermo Fisher Scientific, cat. #11875119), strained using MACS smart strainer 70 μm, and counted on a Countess II cell counter (Thermo Fisher Scientific). Cells were resuspended in Hanks' Balanced Salt Solution (Thermo Fisher Scientific, cat. #88284) to 1,200 cells/μL for immediate downstream processing. Genetically engineered mouse models were generated and barcoded using the Chromium Single Cell 3′ _v3 reagents and workflow (10x Genomics) at Stanford Genome Sequencing Service Center as a commercial service. Libraries were sequenced on Illumina NovaSeq.

Low-pass whole-genome DNA sequencing assay

Genomic DNA was isolated from FFPE samples using PicoPure DNA Extraction Kit (Thermo Fisher Scientific # KIT0103). DNA library construction and sequencing were performed as previously published (7).

Bulk RNA-seq

The bulk RNA sequencing libraries were prepared from dissected areas of FFPE sections for sample sets 1, 2, and 3 as described previously (7) and then prepared for sequencing using the SMART-3SEQ protocol (17). Libraries were sequenced on an Illumina NextSeq 500 instrument using High Output v2.5 Reagent Kit (Illumina # 20024906).

Immunofluorescence on pathology specimens

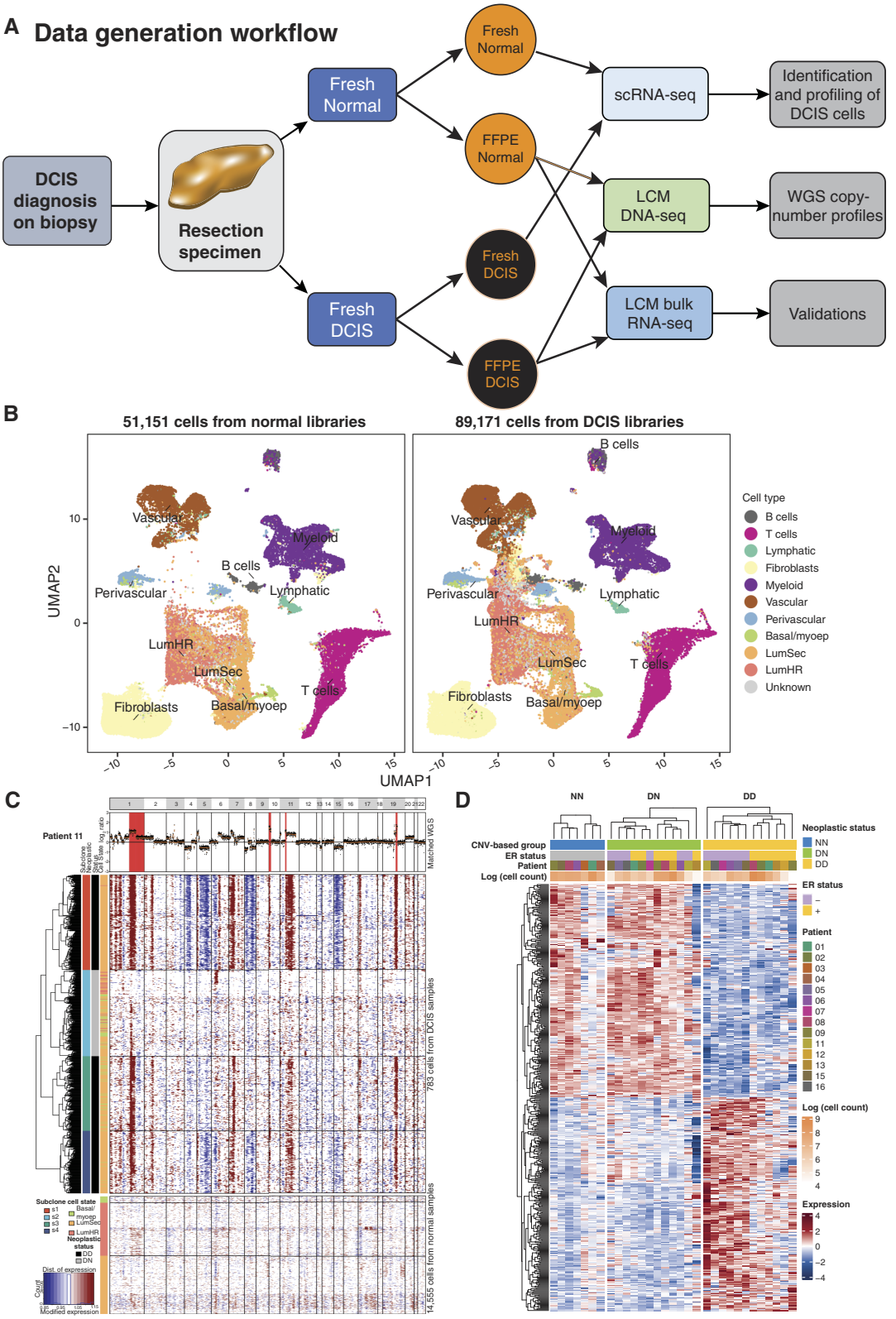
Immunofluorescence (IF) staining was performed on paraffin-embedded tissue microarray sections (4 μm; sample set 4). Briefly, slides were deparaffinized and hydrated using a xylene–ethanol series. Antigen retrieval was carried out at 116°C for 3 minutes in a decloaking chamber in Antigen Unmasking Solution, citrate-based, pH6, at 10 mmol/L (Agilent, cat. #S236984-2) for collagen type IV (Sigma, cat. #AB769; 1:50) and ACTA2 (Thermo Fisher Scientific, cat. #53-9760-82; 1:3,000) detection. For Lamc2 (SantaCruz Biotech, cat. #c-28330; 1:50), proteinase K (Agilent Technologies, cat. #S302080-2) digestion was done. Slides were incubated with the primary antibody at room temperature for 45 minutes For detection, Alexa Fluor–conjugated secondary antibodies at 1:700 dilution were used. Counter staining was performed using ProLong Gold Antifade with DAPI (Thermo Fisher Scientific, cat. #P36935). Slides were scanned using a Leica microscope slide scanner using Ariol Software (Leica Biosystems).

Combined IF and RNA *in situ* hybridization assay

Codetection of ACTA2 protein and *COL4A1* and *LAMC2* RNA were performed using all Bio-Techne/ACD instructions for performing *in situ* hybridization using RNAscope 2.5 HD Detection Kit-RED (cat. #322360) and RNA–Protein Codetection Ancillary Kit (cat. #323180) IHC on FFPE tissue microarray tissue sections (sample set 4). The following RNA probes were used: RNAscope Probe - Hs-COL4A1 (cat. #461881) and RNAscope Probe - Hs-LAMC2 (cat. #501371). IHC for ACTA2 was performed at 1:200 dilution with Abcam, cat. #ab5694. The Cy3 filter was used to detect RNA signals, and the AF647 filter was used to detect ACTA2 protein signal.

BM continuity analysis

To evaluate the continuity of the BM in the IF images, the BM location was annotated using QuPath (18). The percentage of continuity of the BM for each duct was determined by calculating the ratio of the total number of intact pixels to the total number of pixels in the BM. Each duct was identified by a pathologist-annotated mask that indicates the precise location of the duct on the


A
Data generation workflow
DCIS diagnosis on biopsy
Resection specimen
Fresh Normal
Fresh DCIS
Fresh Normal
FFPE Normal
Fresh DCIS
FFPE DCIS
scRNA-seq
LCM DNA-seq
LCM bulk RNA-seq
Identification and profiling of DCIS cells
WGS copy-number profiles
Validations
B
51,151 cells from normal libraries
89,171 cells from DCIS libraries
UMAP1
UMAP2
Vascular
Myeloid
B cells
Perivascular
Lymphatic
LumHR
LumSec
T cells
Fibroblasts
Basal/myoep
Cell type
B cells
T cells
Lymphatic
Fibroblasts
Myeloid
Vascular
Perivascular
Basal/myoep
LumSec
LumHR
Unknown
C
Patient 11
Matched WGS
783 cells from DCIS samples
14,555 cells from normal samples
Subclone cell state
Neoplastic status
Dist of expression
Modified expression
D
NN
DN
DD
CNV-based group
ER status
Patient
Log (cell count)
Neoplastic status
ER status
Patient
Log (cell count)
Expression

core. Calculations were performed for the angles of the tangential line to the surface of the mask for each duct in Python. This process aims to offer the algorithm guidance about the approximate shape of the duct, ensuring a comprehensive scanning of the membrane. With the shape of each duct established, the algorithm evaluates every pixel unit along the membrane to determine whether a particular location is deemed intact. This scanning process was conducted within a confidence range of 250 to 500 pixel units to account for potential variability in annotations.

Whole-genome DNA sequencing data processing and analyses

DNA sequencing data were preprocessed using Burrows-Wheeler Aligner–MEM algorithm v0.7.17 for sequence alignment to the reference genome GRCh38/hg38 and GATK v4.1.7.0 (19–21) to mark duplicates and calibrate reads within the Nextflow-based pipeline Sarek v2.6.1 (22, 23). The recalibrated reads were further processed and filtered for mappability and guanine (G) and cytosine (C) content using the R/Bioconductor package QDNAseq v1.22.0 (24) with R statistical environment v3.6.0. For QDNAseq, 50-kb bin annotations were obtained from QDNAseq.hg38 (v1.0.0, https://github.com/asntech/QDNAseq.hg38). Only autosomal sequences were retained after filtering based on low-depth mappability and GC correction. The R package ACE v1.4.0 (25) was used to estimate and identify the maximum cellularity for different ploidies (2–4). Copy-number aberrations were called using CGHcall v2.48.0 (26).

Bulk RNA-seq data processing

The RNA sequencing libraries were preprocessed using 3SEQtools (17). Read alignment was conducted using STAR v 2.7.3a (27). Gene level reads were obtained using featureCounts Rsubread package v4.0.0 (28). The reference sequence and genomic annotation files were obtained from GENCODE (29).

scRNA-seq data processing and analyses

Preprocessing

To process the raw scRNA-seq data into a cell by gene matrix with read counts, 10X Genomics Cell Ranger v6.0.1 (30) was used. Briefly, the mkfastq module used bcl2fastq to demultiplex the reads. Then the count module aligned and mapped the reads to the human reference genome (reference bundle version "refdata-gex-GRCh38-2020-A") and then estimated read counts for each cell across 36,601 annotated genes. The resulting cell-by-gene count matrices were imported into the R v4.2.2 using the Seurat package v4.3.0.1 (31) to perform quality control (QC) and downstream analyses.

QC

Data quality was assessed in the following ways. Cell quality was assessed by examining the distribution of reads and genes detected per cell and the proportion of reads mapping to mitochondrial genes. A gene was considered to be "detected" if at least one read was mapped to it. Rare gene features were removed from the dataset by excluding genes that were detected in fewer than 10 cells per sample. Mitochondrial genes were identified as features whose corresponding HUGO symbol was prefixed by the string "MT-." At a minimum, cells were assumed to be of low quality and excluded if they had fewer than 200 detected genes or if mitochondrial reads comprised more than 20% of the total reads. In addition, cells with extremely high numbers of reads or genes detected were filtered out based on sample-specific thresholds. The sample-specific thresholds were determined by examining their distributions and removing small shoulders to both sides of the major peak to remove potential dead cells and multiplets.

Eight libraries (three DCIS and five normal) were excluded from downstream analyses due to an insufficient number of high-quality cells (≈1,000 or fewer), and two IBC libraries were excluded due to low sample size. The remaining 20 libraries (7 normal and 13 DCIS) containing 140,322 cells were derived from 14 patients: five patients with paired normal and DCIS libraries (10 libraries), seven patients with DCIS-only libraries (eight libraries, with two DCIS samples containing specimens/libraries from the same patient), and two patients with normal-only patient libraries (two libraries).

Cell integration

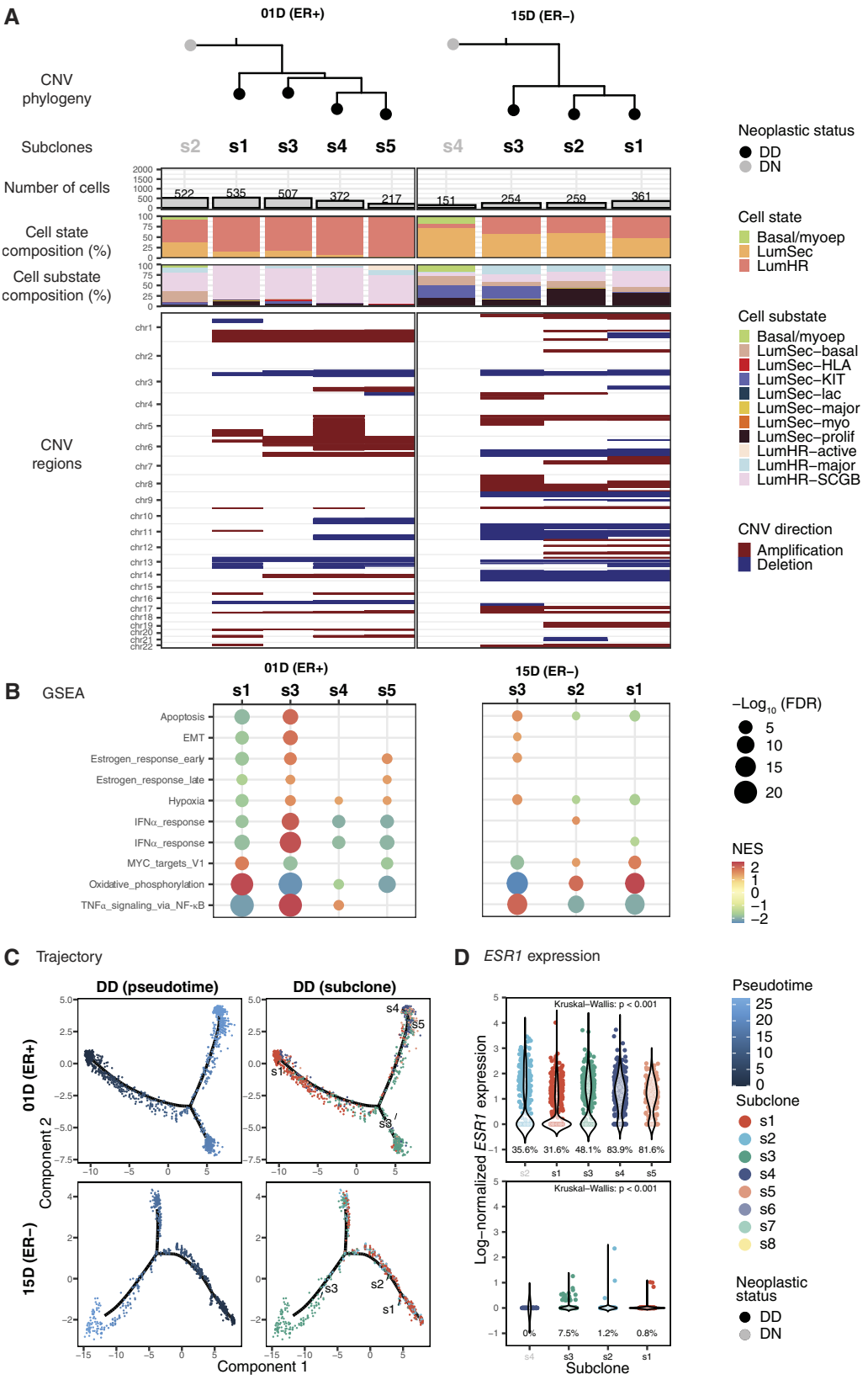
Quality-filtered count matrices from each library were merged into one dataset for downstream analyses and integrated to account for variation among the eight sequencing batches. Integration was performed using Seurat functions (31) that involve (i) merging sample count matrices by batch. In addition (ii), within each batch, counts were log-normalized, and the top 2,000 most variable gene features were selected based on variance-stabilizing transformation. Next, (iii) variable gene features across batches, i.e., integration features, were selected from the union of top batch-wise variable features, and these were used to reduce the dimensionality of each dataset and identify mutual nearest neighbors, i.e., anchors. Integration returned a single matrix of log-normalized and batch-corrected counts for each cell and integration feature; hereafter, "integrated data."

Clustering

To visualize cell expression profiles across all samples, the integrated data were scaled, reduced in dimension using principal component analysis on the top 30 components, and subsequently reduced onto two-dimensional (2D) Uniform Manifold Approximation and Projection (UMAP) coordinates using Seurat

Figure 1.

Overview of workflow. **A,** Laboratory workflow to obtaining multimodal sequencing data from resected specimens (sample set 1) to infer cell phenotypes from scRNA and DNA data and infer cell fractions from specimen-matched bulk RNA data for downstream analyses. LCM, laser capture microdissection. **B,** UMAP embedding plots based on single-cell transcriptome profiles from 140,322 cells across 13 DCIS sample libraries (89,171 cells) and 7 normal breast sample libraries (51,151 cells) obtained from 14 patients. Data integration and dimension reduction was performed on the entire collection of cells to identify 11 primary cell types. After performing the embedding analysis on the latter, the cells were split depending on whether they were derived from a normal breast sample or DCIS sample library. The inferred cell type was used to color each coordinate. For each cell type, the ID is labeled at the median position of all cells in that category. **C,** A representative example of patient tumor-matched copy-number motifs inferred from a low-pass WGS (top) and inferred from single-cell DCIS (middle) libraries. A panel of seven normal libraries was used as the reference (bottom). **D,** Relative expression of pseudobulk samples of epithelial cells from DCIS libraries inferred to be either DD (17,044 cells from 12 pseudobulk samples) or DN (14,309 cells from 12 pseudobulk samples) and cells from normal samples (NN, 14,555 cells from seven pseudobulk samples). The top 250 upregulated and 250 downregulated differentially expressed genes (from DESeq2) ranked by adjusted *P* value between DD and NN are displayed with expression of DN pseudobulk samples shown for comparison. Color bars indicate neoplastic status (DD, DN, or NN), ER status (DD and DN pseudobulk samples only), patient ID, and cell count (natural log scale).


A
01D (ER+)
15D (ER–)
CNV phylogeny
Subclones
s2 s1 s3 s4 s5
s4 s3 s2 s1
Number of cells
522 535 507 372 217
151 254 259 361
Cell state composition (%)
Cell substate composition (%)
CNV regions
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22
Neoplastic status
DD
DN
Cell state
Basal/myoep
LumSec
LumHR
Cell substate
Basal/myoep
LumSec–basal
LumSec–HLA
LumSec–KIT
LumSec–lac
LumSec–major
LumSec–myo
LumSec–prolif
LumHR–active
LumHR–major
LumHR–SCGB
CNV direction
Amplification
Deletion
B GSEA
01D (ER+)
15D (ER–)
s1 s3 s4 s5
s3 s2 s1
Apoptosis
EMT
Estrogen_response_early
Estrogen_response_late
Hypoxia
IFNα_response
IFNα_response
MYC_targets_V1
Oxidative_phosphorylation
TNFα_signaling_via_NF-κB
–Log10 (FDR)
5
10
15
20
NES
2
1
0
–1
–2
C Trajectory
DD (pseudotime)
DD (subclone)
01D (ER+)
15D (ER–)
Component 2
Component 1
s1 s2 s3 s4 s5
Pseudotime
25
20
15
10
5
0
Subclone
s1
s2
s3
s4
s5
s6
s7
s8
Neoplastic status
DD
DN
D ESR1 expression
Kruskal–Wallis: p < 0.001
Log–normalized ESR1 expression
35.6% 31.6% 48.1% 83.9% 81.6%
0% 7.5% 1.2% 0.8%
Subclone

functions (32). Unsupervised clustering of the cells was performed with Seurat functions (32) by constructing a shared nearest-neighbor graph based on the top 30 principal component analysis components and then optimizing the standard modularity function based on the Louvain algorithm with the resolution parameter set to 0.5.

Cell type inference

To infer a cell type classification for each cell based on gene expression, the R package scSorter v0.0.2 (33) was used along with a curated set of signature genes (Supplementary Table S2). The curated signature gene set consisted of genes to profile the 10 major cell types in normal breast tissue, including basal/myoepithelial (basal/myoep) cells, luminal secretory (LumSec) cells, luminal hormone-responsive (LumHR) cells, B cells, fibroblasts, lymphatic cells, myeloid cells, perivascular cells, T cells, and vascular cells (see Supplementary Table S3 from Kumar and colleagues, ref. 34), and 11 epithelial cell substates, including basal/myoep, LumSec-basal, LumSec-HLA, LumSec-KIT, LumSec-lac, LumSec-major, LumSec-myo, LumSec-prolif, LumHR-active, LumHR-major, and LumHR-SCGB (see Supplementary Table S10 from Kumar and colleagues, ref. 34). Specifically, the signature genes for the three major epithelial cell states in the former set were replaced by the signature genes for the 11 epithelial cell substates in the latter set. Cell type inference was performed using the top 2,000 most variable genes, excluding rarely expressed genes that are detected in ≤10% of cells. Counts were log-normalized, and a scSorter (33) tuning parameter α of 0.2 was selected to allow for unknown cells and represent the most common composition of cell types across a range of α values.

MCF10A analysis

The MCF10A cell line scRNA-seq (35) was processed using the same workflow above as sample set 1. Briefly, the raw scRNA-seq count matrix was obtained from GSE200981. The same set of QC metrics used on our study data were applied to filter out low-quality cells. Untreated cells at time point 0 (T0) were subset for dimensionality reduction and UMAP coordinates. Cell type for each cell was inferred using the same method used for our study set but with a reduced cell type set, focusing only on three major cell types, basal/myoep, LumSec, and LumHR, excluding the seven immune cell types.

Copy-number variation inference

We distinguished the DCIS cells in DCIS libraries ("DD," DCIS samples, DCIS cells) from the nonneoplastic epithelial cells in DCIS libraries ("DN," DCIS samples, normal cells) based on their RNA expression–inferred copy-number variation (CNV) profile. The inferCNV package v1.10.1 was used to infer the copy-number states of cells from DCIS samples using scRNA-seq data. To this end, a panel of 51,151 normal epithelial cells from the seven normal scRNA-seq libraries was constructed to estimate the diploid state as a reference. A cutoff of 0.1 for the minimum average read counts per gene among reference cells, default settings for the Hidden Markov Model, and denoising filters were used. In addition, the analysis mode of "subclusters" was used to predict CNV at the levels of subpopulations.

Whole-genome sequencing (WGS) CNV profiles were available for eight of the DCIS samples in the final analysis data set. The CNV profiles were initially assessed through a manual review process based on investigators' knowledge. During this process, specific DNA-inferred CNV regions were identified to serve as decision rules for classifying each individual cell as either a neoplastic or nonneoplastic cell. For a given cell within a scRNA-seq sample, if the number of RNA-inferred CNV regions aligning with the decision rule exceeded a sample-specific threshold, the cell was designated as a neoplastic cell. In cases in which scRNA-seq samples lacked a colocated DNA-inferred CNV profile, the set of decision rule regions was derived from the CNV profile of the scRNA-seq sample itself. In total, 17,044 DD and 14,309 DN cells were identified in 13 DCIS libraries, and 14,555 NN (Normal samples, Normal cells) cells were identified in seven normal libraries. The decision rules and CNV profiles for the DCIS samples are provided in Supplementary Table S3.

DCIS copy-number state phylogenies

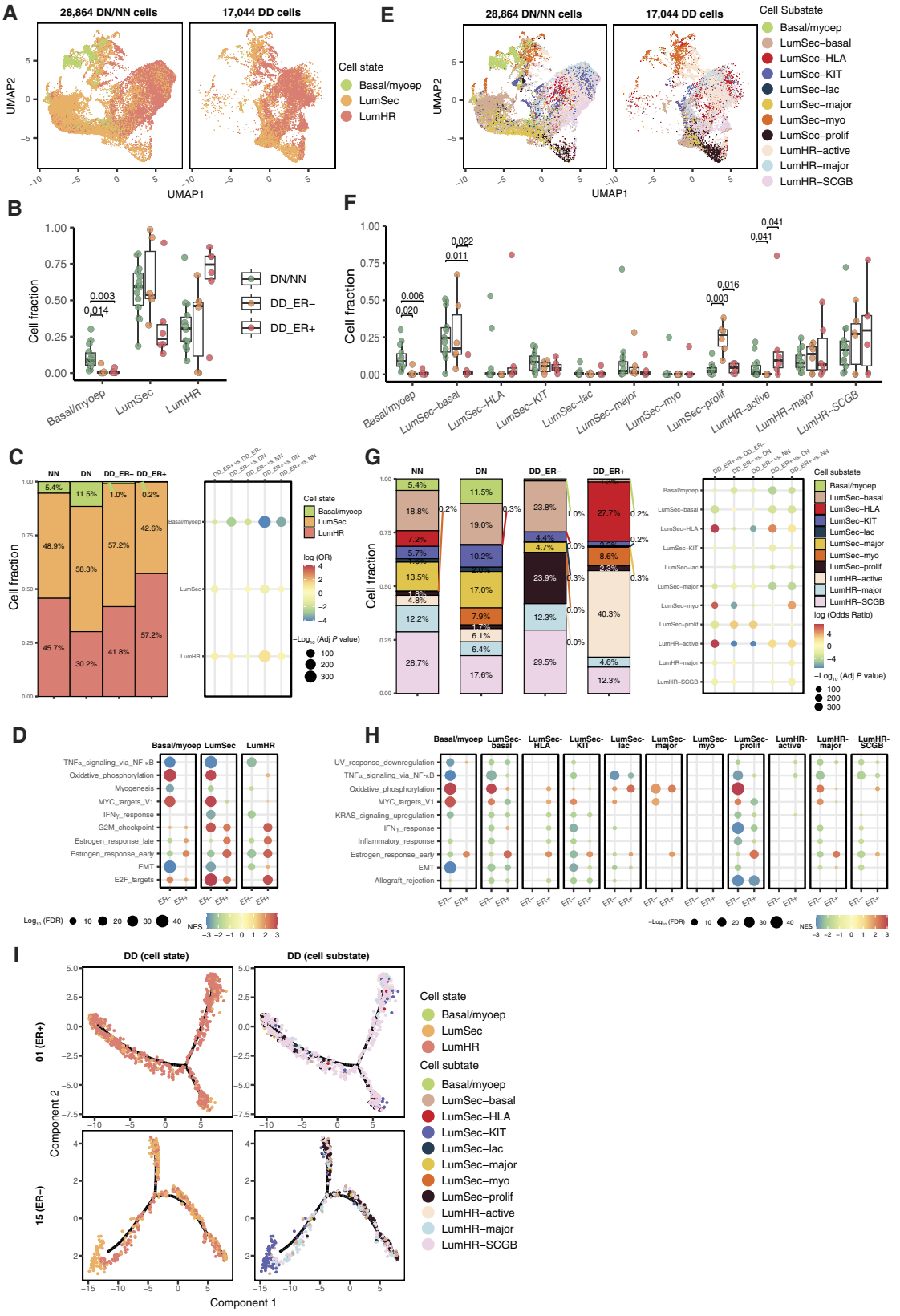
To investigate tumor heterogeneity, we examined cell subpopulations within each DCIS sample with different CNV profiles as identified by Leiden clustering (36) implemented within the inferCNV "subcluster" mode. The cell populations, hereafter "CNV subclones," were characterized by neoplastic status and cell state composition. To show the relationship between CNV subclones and infer alteration histories within a DCIS sample, a phylogenetic tree for each sample was constructed by calculating the Euclidean distance between subclones based on gene-level–predicted CNV states, building a neighbor-joining tree, and then rooting the tree using an outgroup with no copy-number alterations. The R packages ape v5.7.1 (37) and those within treedataverse v0.0.1 were used for tree-building and visualization. For each patient, differential expression (DE) analysis was performed to compare each DD subclone against all the other DD subclones using the FindConservedMarkers function from Seurat package v4.3.0.1 (31) in R. For gene set enrichment analysis (GSEA), all genes were preranked by the log fold change of the average expression from the Seurat results prior to the enrichment analysis of hallmark pathways using the fgsea package v1.24.0. The collection of pathways was acquired from MSigDB (38, 39) via the msigdbr package v7.5.1. All *P* values were adjusted to control for multiple testing using the Benjamini–Hochberg procedure.

Pseudotime trajectory inference

The Monocle package v2.26.0 (40–42) was used to infer the pseudotime trajectory of epithelial cell expression for each patient individually. To prepare the data, the raw read count matrix of the epithelial cells was filtered to exclude rarely expressed genes

Figure 2.

Heterogeneity of CNV profiles in DCIS. **A,** Phylogenetic trees constructed using inferred subclone CNVs show the relationship between subclones that were assigned neoplastic status classifications (DD vs. DN) as shown by tree tips and subclone label colors. The number of cells, cell type compositions (states and substates), and inferred CNV regions by subclone are shown in the bar charts and tile plots. **B,** GSEA results for a given DCIS subclone as compared with others in the same sample are shown. NES, normalized enrichment score. **C,** Subclone states overlayed onto trajectory maps. Cells in the left column are colored by pseudotime, in the right column are colored by subclone states. **D,** Log-normalized expression of *ESR1* gene by subclone states. The percentages above the *x*-axis represent the percentage of cells expressing *ESR1* (log-normalized expression greater than 0) in each subclone. Difference in *ESR1* expression between DD subclones was tested using the Kruskal–Wallis test within each patient.


A
28,864 DN/NN cells
17,044 DD cells
UMAP2
UMAP1
Cell state
Basal/myoep
LumSec
LumHR
E
28,864 DN/NN cells
17,044 DD cells
Cell Substate
Basal/myoep
LumSec–basal
LumSec–HLA
LumSec–KIT
LumSec–lac
LumSec–major
LumSec–myo
LumSec–prolif
LumHR–active
LumHR–major
LumHR–SCGB
B
Cell fraction
DN/NN
DD_ER–
DD_ER+
Basal/myoep
LumSec
LumHR
F
Cell fraction
Basal/myoep
LumSec–basal
LumSec–HLA
LumSec–KIT
LumSec–lac
LumSec–major
LumSec–myo
LumSec–prolif
LumHR–active
LumHR–major
LumHR–SCGB
C
NN
DN
DD_ER–
DD_ER+
Cell fraction
Cell state
Basal/myoep
LumSec
LumHR
log (OR)
–Log10 (Adj P value)
G
NN
DN
DD_ER–
DD_ER+
Cell fraction
Cell substate
log (Odds Ratio)
–Log10 (Adj P value)
D
Basal/myoep
LumSec
LumHR
TNFα_signaling_via_NF-κB
Oxidative_phosphorylation
Myogenesis
MYC_targets_V1
IFNγ_response
G2M_checkpoint
Estrogen_response_late
Estrogen_response_early
EMT
E2F_targets
–Log10 (FDR)
NES
H
Basal/myoep
LumSec-basal
LumSec-HLA
LumSec-KIT
LumSec-lac
LumSec-major
LumSec-myo
LumSec-prolif
LumHR-active
LumHR-major
LumHR-SCGB
UV_response_downregulation
TNFα_signaling_via_NF-κB
Oxidative_phosphorylation
MYC_targets_V1
KRAS_signaling_upregulation
IFNγ_response
Inflammatory_response
Estrogen_response_early
EMT
Allograft_rejection
–Log10 (FDR)
NES
I
DD (cell state)
DD (cell substate)
01 (ER+)
15 (ER–)
Component 2
Component 1
Cell state
Basal/myoep
LumSec
LumHR
Cell subtate
Basal/myoep
LumSec–basal
LumSec–HLA
LumSec–KIT
LumSec–lac
LumSec–major
LumSec–myo
LumSec–prolif
LumHR–active
LumHR–major
LumHR–SCGB

(i.e., genes found in 5% or fewer cells or with an expression value below 0.1). Following Monocle's unsupervised procedure of identifying cell-ordering genes, t-distributed stochastic neighbor embedding (t-SNE) was first applied on the top 15 principal components of the filtered data to project them into two dimensions, and clusters were identified by the densityPeak algorithm (43) to detect genes that differ between the clusters, which were subsequently used as ordering genes. The Discriminative Dimensionality Reduction with Trees (40) algorithm was applied on the filtered data to reconstruct a trajectory, with cells ordered using the top 1,000 significant genes that differ across the Monocle-inferred clusters. The root point of the trajectory was selected arbitrarily by Monocle.

Pseudobulk analysis

To identify differentially expressed genes, raw cell-level counts were summed from each patient, cell state, cell substate, and subclone, and then DE and GSEA analyses were performed for various comparisons. DE analysis was performed using the DESeq2 package v1.38.3 (44) in R. For GSEA, all genes were preranked by the Wald statistics from the DESeq2 results prior to the enrichment analysis of hallmark pathways using the fgsea package v1.24.0. All *P* values were adjusted to control for multiple testing using the Benjamini–Hochberg procedure.

Cell type abundance imputation

The CIBERSORTx tool (45) was used to infer the cell type composition within bulk RNA-seq data sets (sample sets 1–3), using a published single-cell data set (34). For each bulk data set, its raw gene count matrix was used as the input, whereas the external scRNA-seq count matrix with original cell assignments was used to construct the cell type signature matrix. The matrix was randomly downsampled to include a maximum of 3,000 cells per cell type for the purpose of computing efficiency. The fraction of cells with identical identities showing evidence of gene expression was set to 0. To ensure robust analysis, 100 permutations were performed for *P* value calculation. S-mode batch correction, designed specifically for scRNA-seq–derived signature matrices, was applied to correct for cross-platform batch effects. The raw imputed cell fractions were then used to calculate the fraction of each epithelial cell type relative to the total fraction of all epithelial cell types.

Statistical analyses

To compare imputed cell type abundance across sample types in bulk datasets and to compare the cell type fractions across sample types in scRNA-seq data, the Wilcoxon rank-sum test was used. In the case in which bulk data had a patient-matched design (sample set 1), the Wilcoxon signed-rank test was used instead. To account for multiple testing arising from various cell types, the resulting *P* values were corrected using the Benjamini–Hochberg method. The Fisher exact test was used to compare the overall fractions of cells of each contrast group within each cell type. The Pearson correlation test was used to evaluate the correlation between the cell type abundance estimated from the scRNA-seq data and the corresponding bulk data (sample set 1) on a per-sample basis. The Kruskal–Wallis test was used to test whether there are any differences in *ESR1* expression between DD subclones within each patient.

Outcome analysis

Patients in sample set 3 were divided into "high" and "low" cell fraction groups based on the median of each respective cell substate. Associations with time to IBC recurrence were quantified using a Cox proportional hazards model (46). Kaplan–Meier plots as implemented in the R packages survival v3.3-1 and survminer v0.4.9 were used to visualize outcome differences.

In vitro invasion assay

Cell culture

MCF10A human mammary epithelial cells (ATCC) were cultured in DMEM/Nutrient Mixture F-12 (DMEM/F12) medium (Thermo Fisher Scientific) supplemented with 5% horse serum (Thermo Fisher Scientific), 20 ng/mL EGF (Peprotech, Inc.), 0.5 μg/mL hydrocortisone (Sigma), 100 ng/mL cholera toxin (Sigma), 10 μg/mL insulin (Sigma), and 1% penicillin/streptomycin (Thermo Fisher Scientific). Cells were passaged every 3 to 4 days with 0.05% trypsin/EDTA and cultured in a standard humidified incubator at 37°C and 5% CO_2.

Acini formation

Mammary acini using MCF10A mammary epithelial cells were generated as previously described (47). Briefly, single-cell suspensions of MCF10A cells were seeded onto reconstituted BM (rBM) in 2 mL of growth media supplemented with 2% rBM (48). After 4 days, the media was replenished with rBM-supplemented media. On day 5, acini were extracted by treatment with 50 mmol/L EDTA in PBS followed by cell scrapping. After 20 minutes of incubation on ice, the acini-containing EDTA mixture was spun down for 5 minutes at 500 *g* at 4°C, resuspended in the growth media, and centrifuged at 500 *g* for another 5 minutes. After centrifugation, the supernatant was aspirated, and acini were resuspended in DMEM/F12 media.

Alginate preparation and acini encapsulation in hydrogels

High–molecular weight sodium alginate was synthesized and used to develop interpenetrating hydrogels with specific mechanical

Figure 3.

Identification and comparison of epithelial subtypes. **A,** UMAP embedding plots based on 45,908 epithelial cells (left, 28,864 DNNN; right, 17,044 DD) colored by the three major epithelial cell states found in the normal breast. Data integration and dimension reduction was performed on epithelial cells prior to UMAP embedding. **B,** Relative cell fractions of three major epithelial cell states at the pseudobulk level for each sample for DN/NN, DD ER−, and DD ER+. Variation in cell fraction by group was tested using the Wilcoxon rank-sum test, and *P* values were adjusted using the Benjamini–Hochberg method to account for multiple comparisons. The results with an adjusted *P* value of less than 0.05 are shown. **C,** Corresponding relative frequencies for the cell state compositions within DN, NN, DD ER+, and DD ER− groups are summarized as stacked bars, with text showing the percentage of cells. The dot grid shows the resulting odds of cell state membership based on CNV and ER status. The color of the dots indicates the log OR, and the size indicates significance. **D,** Selected significant pathways from GSEA differentially enriched between pseudobulk samples from DN/NN and DD of three major cell states in ER− DCIS and ER+ DCIS, analyzed separately. The color of the dots indicates the normalized enrichment score (NES), and the size indicates significance measured as the negative $\log_{10}$-adjusted *P* value. **E–H** present results at finer resolution, in which 11 epithelial cell substates are considered instead of the three major states. They correspond to panels **A–D**, respectively. **I,** Cell trajectory analysis based on DN/NN, DD ER−, and DD ER+ for selected patients 01 and 15. Cells in the left column are colored by the three major epithelial cell states, and cells in the right column are colored by the 11 epithelial cell substates.

stiffness (2.4 kPa) for acini culture, as described previously (47). Briefly, alginate was first mixed with rBM matrix (Matrigel, Corning) on ice. Next, MCF10A acini were added to this solution. Finally, the polymer solution containing acini was mixed with calcium cross-linker, and the mixture was placed in an incubator to allow the interpenetrating polymer network (IPN) hydrogel to form. After 1 hour, cell culture media were added to the gels.

IF of acini

Acini were fixed in 4% paraformaldehyde in serum-free DMEM/F12 at room temperature. After fixation, acini were washed twice in PBS for 15 minutes and stained with antibodies. The following antibodies were used for detection: Alexa Fluor 488–conjugated anti–laminin-5 antibody, clone D4B5 (Millipore Sigma, MAB19562X, 1:200 dilution), anti–collagen-IV mouse antibody (Sigma, #SAB4500369, 1:200) and Alexa Fluor 690 goat anti-mouse antibody (#A21240, 1:1,000), and nuclear stain Hoechst 33342 (#H3570, 1:1,000).

Hepsin and collagenase IV treatment

Hepsin (R&D Systems #4776SE010) was reconstituted to a stock concentration of 100 μg/mL in an assay buffer as per the manufacturer's instructions. The stock solution was further diluted to 10 μg/mL in MCF10A growth media, and acini were incubated in the diluted solution overnight. Collagenase IV powder (Sigma #17104019) was reconstituted in Hank's Balanced Salt Solution to a stock concentration of 25 U/mL, and acini were treated with a 1:1,000 dilution for 1 hour. For the acini treated with hepsin and collagenase, the acini were initially incubated with hepsin overnight and then treated with collagenase IV for 1 hour. Acini were encapsulated in hydrogels immediately after treatment.

Measurement of COL4 thickness

From the confocal images of COL4, the distance from the inner surface to the outer surface of COL4 was measured at three random locations in the equatorial cross-section, and an average of these measurements was taken to obtain the average thickness for each acinus in the experiment.

Measurement of BM breaching, invasive acini, and circularity

BM breaching was defined by the ratio of the BM-enclosed area to the acini area in the plane of invasion, with lower ratio corresponding to increased breaching. The BM-enclosed area and the acini area in the plane of invasion were computed in ImageJ by drawing a manual outline around the BM and the acini, respectively. Invasive acini were identified manually as those structures in which a group of cells had migrated out of the originally circular footprint of the acini in a 2D field of view. The percentage was calculated based on the number of such invasive acini compared with the total number of acini in a field of view. The acini invasion was further quantified by measuring the acini circularity in ImageJ, in which decreased circularity corresponds to increased invasion.

Confocal microscopy

Microscope imaging was performed using a laser-scanning Leica SP8 confocal microscope or a Nikon Ti2-E inverted microscope, both fitted with a temperature and incubator control suitable for live imaging (37°C, 5% CO_2). For capturing fluorescent images, the Leica microscope used a 25× NA 0.91 water objective. For capturing phase-contrast images of acini, the Nikon microscope used a 10× NA 0.45 dry objective.

Data availability

The datasets generated in this study, namely sample sets 1 and 2, have been deposited in the HTAN database (https://humantumoratlas.org) under the HTAN study phs002371. The publicly available dataset, sample set 3, is also available in this repository. Access to the raw sequencing data requires database of Genotypes and Phenotypes (dbGaP) approval, which can be requested at the dbGaP study page, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002371. Dataset mapping IDs are provided in Supplementary Table S1. Publicly available data generated by others used by the authors include the following: (i) hallmark pathway gene sets that were acquired from MSigDB (38, 39) via the msigdbr package v7.5.1, (ii) the processed scRNA-seq data object for all cells from Kumar and colleagues (34), which was obtained from CZ CELLxGENE Discover at https://cellxgene.cziscience.com/collections/4195ab4c-20bd-4cd3-8b3d-65601277e731, and (iii) the processed scRNA-seq data matrix for the MCF10A cell line from Paul and colleagues (35), which was obtained from the Gene Expression Omnibus database at GSE200981. All other raw data are available upon request from the corresponding author.

The scripts to reproduce the analyses presented in this article are available publicly through a GitLab repository (https://gitlab.oit.duke.edu/dcibioinformatics/pubs/pca-dcis-scrna-seq).

Results

scRNA-seq distinguishes normal from neoplastic epithelial cells

We performed scRNA-seq on DCIS ($n = 16$) and matched synchronous normal tissues ($n = 12$) from 15 patients undergoing mastectomy (sample set 1). We excluded three DCIS and five normal samples based on QC metrics. The QC results for all samples are shown in Supplementary Fig. S1, and Supplementary Table S1 lists the 13 DCIS and 7 normal samples from 14 patients that passed QC with patient demographic features. Both the DCIS and normal specimens were bisected, with one half used for single-cell dissociation and scRNA-seq, and the other half was FFPE to confirm the presence and extent of DCIS and normal breast epithelium (**Fig. 1A**). From the FFPE samples, we performed WGS and bulk RNA-seq on microdissected areas of histologically confirmed DCIS.

Two-dimensional UMAP embedding plots of cells from DCIS and matched normal tissue samples colored by cell type (**Fig. 1B**) show that the overall architecture of the clustering was preserved between the two sample types but that some clusters were only present in DCIS samples from specific patients (Supplementary Fig. S2A). Based on histologic examination of the facing FFPE section, we recognized that the dissociated cells from DCIS samples were comprised of a mixture of neoplastic and nonneoplastic epithelial cells (Supplementary Fig. S2B). We used the scRNA expression to infer cell-specific genomic CNV profiles and compared those with WGS-based CNV profiles derived from microdissected DCIS from the adjacent paraffin block (**Fig. 1C**; Supplementary Fig. S2C). Epithelial cells with CNVs from DCIS specimens were categorized as DCIS cells ("DD") and those lacking these CNVs were categorized as "DN." Additionally, because our study included a series of normal or uninvolved specimens, these epithelial cells were categorized as "NN." To further validate that our inferred CNV-based approach distinguished DD from DN epithelial cells, we compared aggregated gene expression (pseudobulk samples) between NN and DD cell populations. A heatmap of the top 250 significantly upregulated and 250 downregulated genes comparing the DCIS

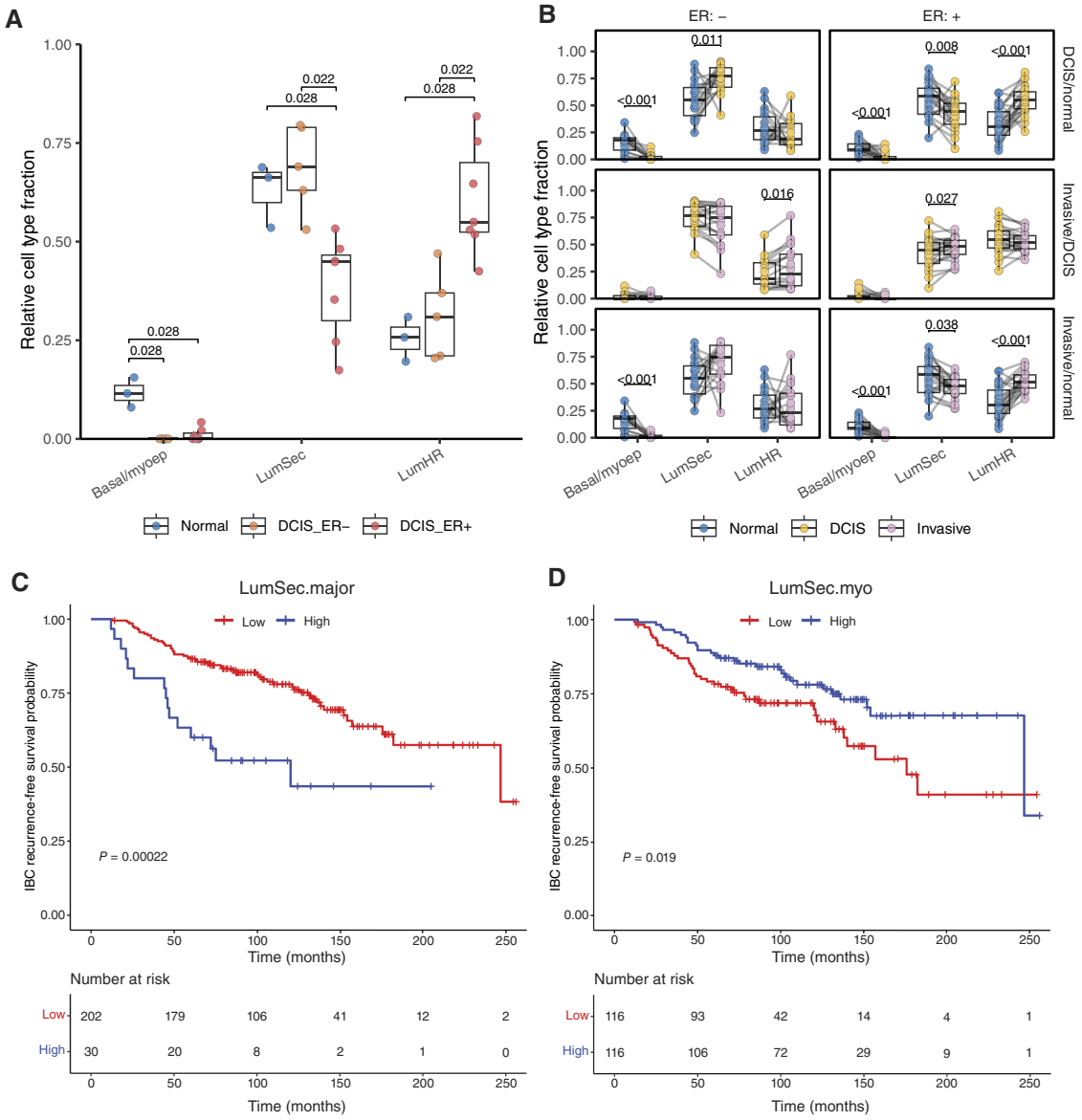
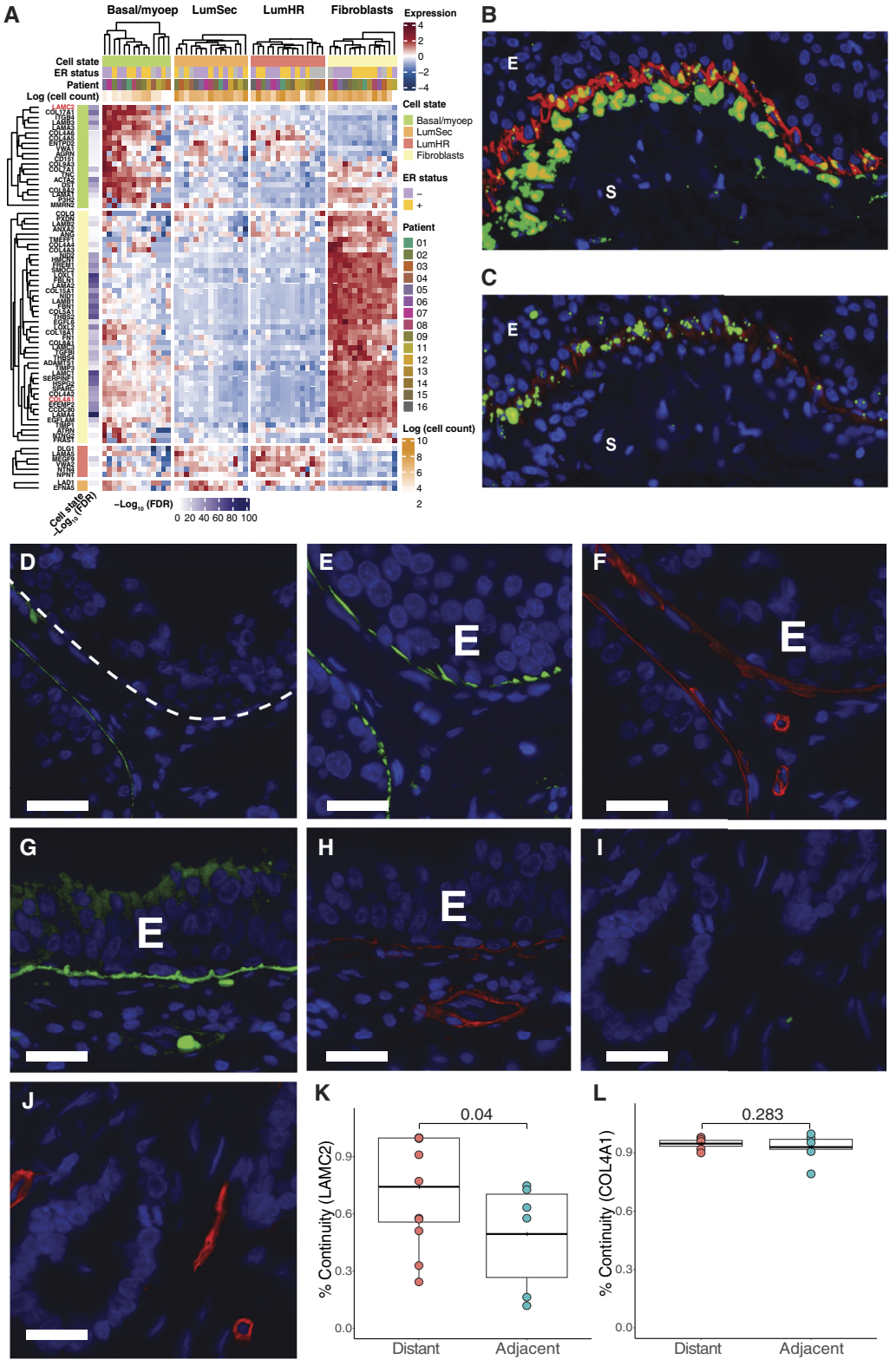


Figure 4.

Analysis of cell type proportions in independent DCIS cohorts. **A,** Analysis of imputed cell fractions based on three major epithelial cell states found in the normal breast from bulk RNA-seq datasets of the FFPE sample set matching the scRNA-seq samples as depicted in **Fig. 1A** with three normal, seven ER+ DCIS, and five ER− DCIS libraries obtained from 11 patients (sample set 1). Differences were tested using the unpaired two-sample Wilcoxon rank-sum test. *P* values were adjusted using the Benjamini–Hochberg method to account for nine comparisons. The results with an adjusted *P* value of less than 0.05 are shown. **B,** Analysis of imputed cell fractions based on three major epithelial cell types found in normal breast from bulk RNA-seq datasets of 126 patient-matched normal breast, DCIS, and IBC libraries (24 ER+ DCIS, 18 ER− DCIS, 42 matched normal, and 42 matched IBC; sample set 2). Points for each patient are connected with a gray line. Pairwise differences were tested using one-sample Wilcoxon signed-rank tests. *P* values were adjusted using the Benjamini–Hochberg method to account for 18 comparisons. **C** and **D,** Kaplan–Meier curves of time to IBC recurrence for high and low levels of LumSec-major (**C**) and LumSec-myo (**D**) from the analysis of imputed cell fraction bulk RNA-seq datasets of retrospective DCIS case–control cohorts (sample set 3). The *P* values were derived from Cox proportional hazards models to assess the difference in recurrence risk between the two groups. The table below each plot shows the number of patients still at risk of recurrence at each time point.

(DD) with normal cells (NN) is shown in **Fig. 1D** (the complete list of differentially expressed genes is provided in Supplementary Table S4). DN cells were not used in the DE analysis for this heatmap but clustered closely with cells from normal breast (NN; **Fig. 1D**), indicating that the DN cells are admixed normal breast epithelial cells known to exist in these specimens based on histologic examination of the facing block. Comparison of DD with either DN or NN cells revealed similar patterns in significant pathway alterations, including cell-cycle progression and metabolic pathways elevated in DD cells, further supporting the accuracy of the cell level categorization based on inferred copy number (Supplementary Fig. S2D).


A
Basal/myoep
LumSec
LumHR
Fibroblasts
Expression
Cell state
ER status
Patient
Log (cell count)
Basal/myoep
LumSec
LumHR
Fibroblasts
ER status
Patient
Log (cell count)
−Log10 (FDR)
0 20 40 60 80 100
B
E
S
C
E
S
D
E
F
E
G
E
H
E
I
J
K
0.04
% Continuity (LAMC2)
Distant
Adjacent
L
0.283
% Continuity (COL4A1)
Distant
Adjacent

Inferred CNV subclones demonstrate intratumoral genetic and expression heterogeneity

CNV analysis revealed that most of the DD cell populations contained differing copy-number states indicative of separate clones. The polyclonal nature of DCIS has been established (9), and our single-cell analyses allowed us to explore the relationship of these clones with their phenotypic properties, including epithelial cell state. We sought to identify these epithelial cell states using expression signatures based on the recent compendium of single-cell data (34) from more than 200 normal breast samples, classifying the epithelium into three primary epithelial "states" (basal, LumSec, and LumHR) and 11 epithelial "substates" (one basal substate, seven LumSec substates, and three LumHR substates). Based on the expression of characteristic markers, myoepithelial cells comprise the majority of the basal cell state and substate and are hereafter referred to as basal/myoep cells.

Two sample phylogenetic trees with cell states derived from scRNA data are shown in **Fig. 2A** (other cases are shown in Supplementary Fig. S3A). Sample 01D is an estrogen receptor–positive (ER+) DCIS that demonstrates four subclones or branches of DCIS cells (and one branch of nonneoplastic DN cells), and sample 15D is an ER-negative (ER−) DCIS with three neoplastic clones. Fractional breast cell state composition of each of these branches indicates that none of these subclones are a monolithic population at the phenotypic level but that each subclone contains varying proportions of cell states and substates.

In considering the subclones as single populations, each subclone exhibited enrichment of unique combinations of cancer hallmark pathways (**Fig. 2B**; Supplementary Fig. S3B). We analyzed the subclones using a trajectory analysis derived for each patient (**Fig. 2C**; Supplementary Fig. S3C). Subclones are enriched for positions on the individual trajectory map, but there is still substantial scattering of the subclones across inferred phenotypic trajectories (i.e., through pseudotime), indicating the extent of phenotypic differences among patients and within subclones in a given patient. Trajectory maps of nonneoplastic epithelial cells also vary among patients. In some cases, there is good enrichment for subclones on the trajectory maps. Furthermore, among subclones, we observed highly significant pathway differences that are not readily evident either by examining cell state or composite epithelial UMAP position. For example, subclones of 01D exhibit similar cell state distributions, but subclones S1 and S3 have several pathways (e.g., oxidative phosphorylation and TNFα signaling) that distinguish these two populations. Examination of individual cases reveals the complex genetic and phenotypic heterogeneity and indicates that clonal populations are themselves diverse mixtures of different mammary cell states.

ER expression is not uniform across all cells within a DCIS sample. Based on the expression data, we analyzed the level of ER expression across different subclones for each patient (**Fig. 2D**; Supplementary Fig. S3D). We observe differences in ER positivity between subclones in both ER+ and ER− DCIS.

ER status distinguishes DCIS cell state composition

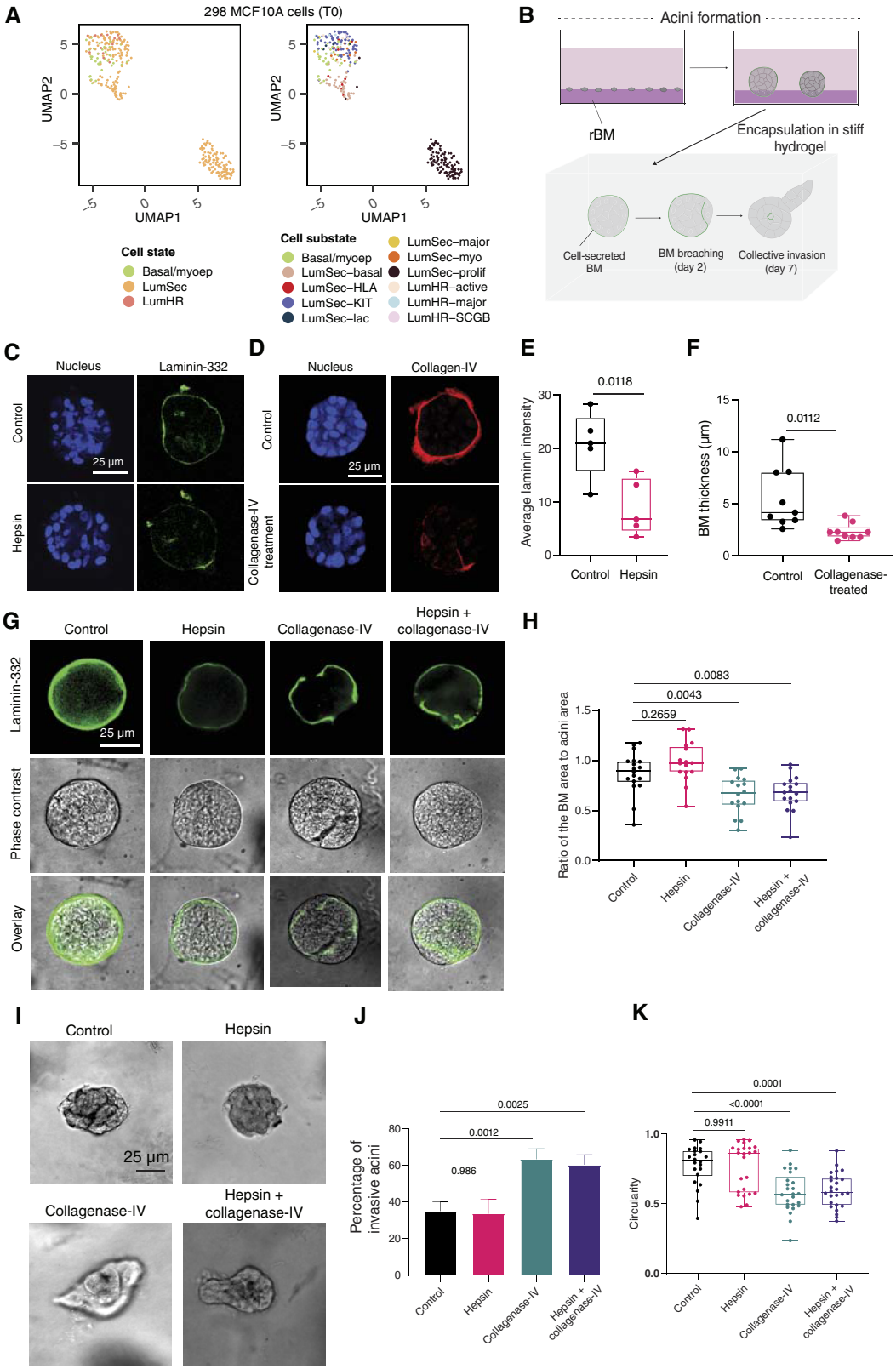
We next analyzed our single-cell sequencing data aggregated by patient, neoplastic status, and ER status. From among 140,322 cells that passed final QC metrics for scRNA-seq, we classified 45,908 epithelial cells from DCIS and normal libraries into three primary states and 11 substates as described above (Supplementary Table S2; ref. 34). The three primary cell states are displayed in epithelial-specific UMAPs separated into DCIS (DD) and non-DCIS cells (DN/NN; **Fig. 3A**). We observed good concordance between the positions of the cells on the UMAP and the three cell states (basal/myoep, LumSec, and LumHR) assigned independently based on sets of signature genes (34). Comparing DCIS (DD) with normal (DN/NN), we observed an overall loss of basal/myoep cells and LumSec cells in the DCIS samples compared with the normal samples.

We analyzed this further at the patient level and confirmed both a lower proportion of basal/myoep cells in both ER+ and ER− DCIS compared with normal samples (**Fig. 3B**, adj. $P = 0.003$ and 0.014, respectively). LumHR cells comprised the greatest proportion of the DD populations from ER+ DCIS, and concomitantly, there were fewer LumSec and basal/myoep cells. Interestingly, ER− DCIS cells exhibited higher proportions of LumSec fractions compared with the ER+ DCIS cells. Aggregating the cell data confirmed highly significant distributional differences between normal (NN or DN), ER+ DCIS, and ER− DCIS cells (**Fig. 3C**). Comparing hallmark pathways within cell states between the normal cells (DN/NN) and ER+ and ER− DCIS cells (DD) revealed substantial differences, with pathways related to epithelial–mesenchymal transition (EMT), proliferation, and estrogen differentially enriched within the cell states between these cell populations (**Fig. 3D**; Supplementary Fig. S4A; Supplementary Table S5).

We next repeated this analysis by applying the more granular 11 epithelial substates (UMAP in **Fig. 3E**). Comparisons between normal, ER+ DCIS, and ER− DCIS cell populations demonstrated significant differences in cell substate compositions and expression pathways (**Fig. 3F–H**). The relative proportions of both the LumSec-basal and LumSec-prolif substates were higher in ER− DCIS compared with ER+ DCIS (adj. $P = 0.022$ and 0.016, respectively). With respect to the LumHR substates, the relative proportion of the LumHR-active substate was significantly lower in ER− compared with ER+ DCIS and normal cells (adj. $P = 0.041$ and 0.041, respectively). ER− DCIS contained substantial proportions of the various LumHR subtypes (i.e., LumHR-major and LumHR-SCGB). Therefore, even though the ER− DCIS cells did not express appreciable levels of the ER, a proportion of

Figure 5.

BM expression in DCIS. **A,** Relative expression of BM genes across the three major cell states and fibroblasts within each pseudobulk samples from scRNA-seq data (sample set 1). Displayed genes were most significantly upregulated (adj. $P < 0.01$) in each specific cell type compared with all other cell types. **B,** Detection of *COL4A1* RNA (green) in DCIS epithelium (E) and adjacent stroma (S), with ACTA2 protein (red) expression demarcating the basal zone of the DCIS sample. **C,** Detection of *LAMC2* RNA (green) in DCIS epithelium (E) and adjacent stroma (S), with ACTA2 protein (red) expression demarcating the basal zone of the DCIS sample. **D,** IF of LAMC2 in DCIS. The white dashed line indicates the edge of the DCIS. E, located adjacent to invasion (sample set 4). **E,** IF of ACTC2 in DCIS (E) located adjacent to microinvasion. **F,** IF of COL4A1 in DCIS (E) located adjacent to invasion. **G,** IF of LAMC2 in DCIS (E) located distant to invasion. **H,** IF of COL4A1 in DCIS (E) located distant to invasion. **I,** IF of LAMC2 in IBC. **J,** IF of COL4A1 in IBC. **K,** Comparison of percentage of continuity for LAMC2 in DCIS ducts between DCIS adjacent to invasion or distant to invasion. **L,** Comparison of percentage of continuity for COL4A1 in DCIS ducts between DCIS adjacent to invasion or distant to invasion. Scale bar, 20 μm.


A
298 MCF10A cells (T0)
UMAP2
UMAP1
Cell state
Basal/myoep
LumSec
LumHR
Cell substate
Basal/myoep
LumSec–basal
LumSec–HLA
LumSec–KIT
LumSec–lac
LumSec–major
LumSec–myo
LumSec–prolif
LumHR–active
LumHR–major
LumHR–SCGB
B
Acini formation
rBM
Encapsulation in stiff hydrogel
Cell-secreted BM
BM breaching (day 2)
Collective invasion (day 7)
C
Nucleus
Laminin-332
Control
Hepsin
25 μm
D
Nucleus
Collagen-IV
Control
Collagenase-IV treatment
25 μm
E
0.0118
Average laminin intensity
Control
Hepsin
F
0.0112
BM thickness (μm)
Control
Collagenase-treated
G
Control
Hepsin
Collagenase-IV
Hepsin + collagenase-IV
Laminin-332
Phase contrast
Overlay
25 μm
H
0.0083
0.0043
0.2659
Ratio of the BM area to acini area
Control
Hepsin
Collagenase-IV
Hepsin + collagenase-IV
I
Control
Hepsin
Collagenase-IV
Hepsin + collagenase-IV
25 μm
J
0.0025
0.0012
0.986
Percentage of invasive acini
Control
Hepsin
Collagenase-IV
Hepsin + collagenase-IV
K
0.0001
<0.0001
0.9911
Circularity
Control
Hepsin
Collagenase-IV
Hepsin + collagenase-IV

these cells retained expression profiles that classified them as LumHR.

Pathway analysis comparing ER+ and ER− DCIS with normal cells demonstrated global differences in TNF signaling (lower in ER− DCIS) and estrogen response (higher in ER+ DCIS) across multiple cell substates (**Fig. 3H**; Supplementary Fig. S4B; Supplementary Table S6). We also noted some significant substate-specific differences between ER+ and ER− DCIS cells, compared with normal cells, including lower expression of the EMT pathway in the basal/myoep substate as well as lower IFNγ response in the LumSec-prolif substate in ER− DCIS compared with ER+ DCIS.

Next, we analyzed the distribution of differentiated states within a given patient sample by overlaying them on the same trajectory analyses shown in **Fig. 2C** (**Fig. 3I**; Supplementary Fig. S3C). In ER+ DCIS, the cell states and substates are distributed throughout the phenotypic trajectory. In ER− DCIS, there is an enrichment of the LumSec-Kit at one end of the trajectory, consistent with the possibility that these cells have stem cell properties. Examination of the other cases (e.g., patients 02 and 15), when LumSec-KIT cells are identified, regardless of ER status, showed that they tend to occupy one end of the phenotypic trajectory (Supplementary Fig. S3C). The shape and distribution of the cell states of the trajectory maps of normal epithelial cells vary between patients, consistent with the findings of considerable cell composition heterogeneity (34). Distribution of the cells on the composite epithelial UMAP (**Fig. 3A**) is an unsupervised measure of cell state. We map each of the CNV-based subclones (**Fig. 2**) for each individual patient (Supplementary Fig. S4C), further highlighting the expression heterogeneity within each subclone.

Cell state analysis in independent DCIS cohorts indicates outcome differences

To confirm and extend these findings, we used deconvolution to estimate the cell state composition from two additional data sets (sample sets 2 and 3) of bulk RNA-seq derived from archival specimens. Sample set 2 is RNA-seq data from laser capture–microdissected epithelium of synchronous samples of normal, DCIS, and IBC cells from the same patient. Sample set 3 is RNA-seq data from macrodissected specimens from two previously reported longitudinal cohorts of DCIS with known disease outcomes (Supplementary Table S1; ref. 7). We first confirmed the validity of the deconvolution method by comparing cell composition from bulk RNA-seq (**Fig. 4A**) with the matched scRNA-seq results (**Fig. 3B**). Correlating with the findings from scRNA-seq, we found lower relative proportion of basal/myoep cells in both ER+ and ER− DCIS vs normal samples (adj. $P = 0.028$, 0.028) and differences between ER− and ER+ DCIS in LumSec (adj. $P = 0.022$) and LumHR (adj. $P = 0.022$). At the individual sample level, we observed significant correlation between scRNA-seq and FFPE bulk RNA-seq data for all three epithelial cell states (Supplementary Fig. S5A).

In sample set 2 (synchronous), the relative proportion of basal/myoep cells was again significantly lower in DCIS versus normal breast for both ER+ and ER− DCIS (adj. $P < 0.001$, **Fig. 4B**). In addition, LumHR cells were significantly enriched in ER+ DCIS compared with the matched normal samples (adj. $P < 0.001$), but this was not observed for ER− DCIS. Conversely LumSec cells were significantly less abundant in ER+ DCIS compared with the matched normal samples (adj. $P = 0.008$) but increased in ER− DCIS (adj. $P = 0.011$). Comparing the matched and synchronous DCIS with IBC, we observed minor increases of LumSec in ER+ IBC and LumHR in ER− IBC. Overall, differences in cell state distribution between synchronous DCIS and IBC are much less pronounced than between normal and DCIS samples.

To investigate the association between the relative abundance of the cell states and substates and risk of invasive progression in DCIS, we analyzed the imputed cell fractions in retrospective DCIS case–control cohorts of patients with a primary diagnosis of DCIS who later either did or did not have IBC progression (sample set 3, longitudinal follow-up). We found that patients with high levels of LumSec-major cells ($P = 0.0002$) and low levels of LumSec-myo cells ($P = 0.019$) were associated with shorter time to IBC recurrence (**Fig. 4C** and **D**; Supplementary Fig. S5B). Other cell states and substates were not significantly associated with invasive progression.

BM gene expression in DCIS

One of the pathways that was sensitive to DCIS subclone progression is the EMT pathway (**Figs. 2B**, **3D** and **H**; Supplementary Fig. S3B), which contains a number of BM-related genes. BM maintenance in the breast is an ongoing and active process (49), and BM loss leads to failure of developmental epithelial structures (50–52). Experimental studies have shown that destruction of the BM is associated with genetic instability and mammary tumorigenesis (53). Thus, we next evaluated whether differences in epithelial state composition could influence the production and integrity of the BM. In our scRNA-seq data, we examined the expression of canonical BM genes (GO:0005604; ref. 54) within the epithelial cell states and fibroblasts (**Fig. 5A**). As expected, fibroblasts demonstrated relatively high expression of many of these genes. Notably, basal/myoep cells also demonstrated significant expression of a set of BM genes, indicating that this epithelial cell state contributes to the ongoing synthesis of the BM. Although there was overlap, expression of some BM genes was uniquely enriched in each cell state, such as *LAMC2* in basal/myoep cells. Among the

Figure 6.

The effects of degrading laminin and COL4 in mammary epithelial cells in a 3D *in vitro* culture model. **A,** UMAP embedding plots based on 298 untreated MCF10A cells, colored by the three major epithelial cell states (top) and the 11 epithelial cell substates (bottom). **B,** Diagram of experimental model of BM invasion of MCF10A acini. **C,** Confocal IF images of nuclei and laminin-332 in control and hepsin-treated acini. **D,** Confocal IF images of nuclei and COL4 in control and collagenase-treated acini. **E,** Average laminin intensity (arbitrary units) in control and hepsin-treated conditions. Each dot corresponds to an acinus ($n \geq 4$ acini). **F,** Average COL4 thickness in control and collagenase-treated conditions. Each dot corresponds to an acinus ($n \geq 9$ acini). **G,** Day 2 analysis of acini BM breaching in the top row. Laminin-332 IF used to measure the location of the BM. Phase-contrast images of acini on day 2 show the boundary of the acini in the middle row. These images are overlayed on the bottom row. **H,** Ratio of the BM area based on laminin stain to the acini area based on phase image. **I,** Phase-contrast images of treated and untreated acini on day 7. **J,** Percentage of invasive acini in control and treated conditions ($n \geq 2$ different gels). **K,** Circularity of acini in control and treated conditions. Each dot corresponds to an acinus ($n \geq 18$ acini). For **E** and **F**, two-sided *t* test with the Welch correction was used. For **H**, **J** and **K**, one-way ANOVA with the Dunnett correction was used.

11 substates, several had prominent BM gene expression, including LumSec-basal and LumSec-KIT (Supplementary Fig. S6A).

In situ analysis of protein and RNA expression of selected BM genes in DCIS samples provided additional evidence that basal/myoep cells contribute to the formation and maintenance of the BM (**Fig. 5B** and **C**). *COL4A1* RNA expression was found in fibroblasts adjacent to the DCIS epithelial compartment (**Fig. 5B**). However, substantial expression of *COL4A1* was also evident in a subset of the ACTA2 (smooth muscle actin)-positive cells (i.e., basal/myoep cells) at the edge of the epithelial compartment. *LAMC2 RNA* was expressed in basal/myoep cells but not fibroblasts (**Fig. 5C**). Based on these data, the epithelial compartment seems to play an active role in the maintenance of the BM.

We further investigated whether BM integrity is associated with breast cancer invasion by examining the physical continuity of specific BM components in a series of DCIS adjacent to areas of microinvasion (defined as ≤1 mm invasive component, sample set 4, $n = 13$). LAMC2 expression was greatly decreased and sometimes absent in some ducts adjacent to the invasion (**Fig. 5D**) despite the presence of basal/myoep cells as defined by ACTA2 expression (**Fig. 5E**). Conversely, COL4A1 protein was found in a consistent pattern around the ducts adjacent to the invasive component (**Fig. 5F**). In contrast, DCIS distant from the invasive component had more continuous expression of LAMC2 and COL4A1 (**Fig. 5G** and **H**). The BM distribution of COL4A1 and LAMC2 is lost in the invasive component of all samples (**Fig. 5I** and **J**). Quantitative measurements of LAMC2 continuity demonstrated a significantly lower percentage of continuity comparing DCIS adjacent to IBC versus DCIS distant ($P = 0.04$), whereas COL4A1 did not show this difference (**Fig. 5K** and **L**). Normal breast demonstrated a continuous COL4A1 and LAMC2 layer surrounding all identified ducts and lobules (Supplementary Fig. S6B and S6C). This shows that DCIS can exist without an intact LAMC2 layer and that loss of this BM component is associated with DCIS near microinvasion, suggesting that this may be an early step in progression. Finally, COL4A1 is lost at the invasive step.

To test the role of laminin and COL4 protein in regulating the invasive phenotype, we used our previously established 3D *in vitro* culture model of BM invasion (47). In this model, MCF10A cells are first cultured in the rBM matrix so that they form organotypic acinar structures that contain a lumen, with the cells exhibiting apicobasal polarity. Importantly, the acini form an endogenous BM around themselves with a layer of laminin-332 on the inside and COL4 on the outside, with the layer of laminin-332 exhibiting a thickness on the range of that observed in DCIS. The acini are then encapsulated in interpenetrating network hydrogels of the rBM matrix and alginate that exhibit an elastic modulus (i.e., stiffness) of 2.4 kPa, similar to what is observed in IBC (47). In this model, the increased stiffness relative to normal breast tissue (100–1,000 Pa) promotes BM invasion by the acini. When MCF10As are cultured in 3D, in interpenetrating network hydrogels, increased stiffness promotes a set of changes in gene expression of MCF10As, mirroring the changes observed in DCIS versus normal patient samples (55). MCF10A cells display a range of cell states and substates that are comparable with the distribution of the normal breast (UMAP of MCF10A grown on plastic; **Fig. 6A**; ref. 35).

In this model of BM invasion (**Fig. 6B**), we examined the impact of perturbing the BM on invasion. In control conditions, the acini exhibit a continuous BM at the periphery, including layers of laminin-332 and COL4 (**Fig. 6C** and **D**). To perturb the BM, the acini were treated with hepsin (56) and collagenase IV to degrade laminin-332 and COL4, respectively (**Fig. 6C** and **D**), which decreased the laminin-332 intensity and the COL4A1 thickness, respectively (**Fig. 6E** and **F**). Next, the acini treated with hepsin, collagenase, both, or none were encapsulated in the interpenetrating network hydrogels and monitored for 7 days. On day 2, whereas the control and hepsin-treated acini did not show breaching of their BM, the acini treated with collagenase (with or without hepsin) breached their BM, as quantified by the ratio of the BM to acini area (**Fig. 6G** and **H**). On day 7, collagenase-treated acini (with or without hepsin) showed higher levels of invasion and decreased circularity than the control and hepsin-treated acini (**Fig. 6I–K**). Together, these data indicate that loss of COL4, but not laminin-332, facilitates higher levels of invasion into the surrounding matrix, in agreement with our *in vivo* data that showed loss of COL4 as always lost at invasion (**Fig. 5**).

A conceptual model for compromise of epithelial integrity

A conceptual model of cancer progression that incorporates the consequences of decreased basal/myoep cell proportions in DCIS is presented in **Fig. 7**. We hypothesized that the relative expansion of neoplastic luminal cells leads to an alteration of the epithelial microenvironment. The relative reduction of cell states (basal/myoep and LumSec-basal) decreases the contributions made by the epithelium to the BM. This results in a decrease in the structural integrity of the DCIS-involved duct. This loss of integrity results in cells losing contact

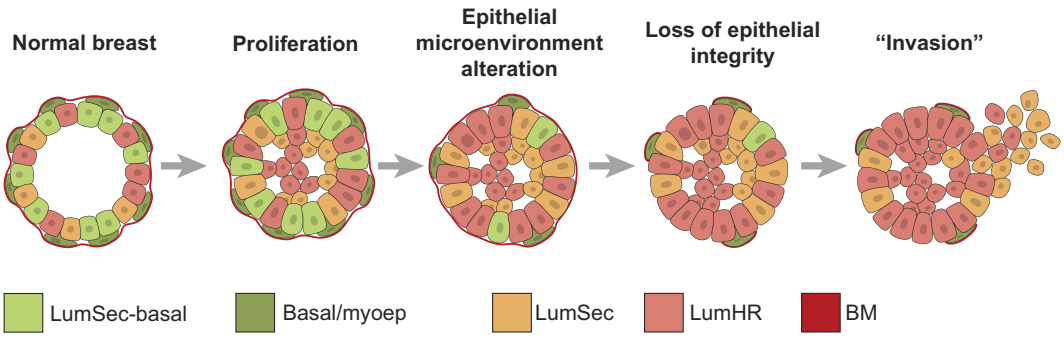


Figure 7.
Conceptual model of DCIS invasion. Hypothesized changes in epithelial cell fractions during disease progression from normal breast to early and late DCIS stages.

with the epithelium compartment and entering the stromal compartment. This passive, loss-of-function event is interpreted as invasion.

Discussion

On the basis of scRNA-seq data, we show that DCIS is comprised of multiple subclones with significant gene expression differences. Furthermore, the DCIS subclones are themselves not monolithic based on their composition of mammary cell states. Normal mammary ducts and lobules are composed of several spatially organized cell types that exist along a differentiation spectrum starting from pluripotent stem cells (57), and a defining feature of DCIS is that it recapitulates normal breast ductal growth patterns, strongly suggesting that the cells retain the ability to differentiate along similar lines. For this reason, in addition to identifying subclones, we also classified the DCIS epithelial cells into phenotypic states according to the recently published single-cell atlas of normal human breast cells, comprising 200 normal breast tissue samples (34). This resource defines 3 states and 11 substates of epithelial cells that we used here and demonstrated that genetic subclones of DCIS contain multiple cell states and that each cell state can be made up of multiple genetic clones.

A key feature of the current study is the discrimination of epithelial cells from the DCIS specimens as "contaminating" normal epithelial cells versus those that are part of the cancer. The use of inferred copy number based on gene expression is a validated approach for single-cell data, and our bulk WGS from the same specimens confirms the presence of the CNVs that were used to distinguish neoplastic from normal cells. Based on histology and inferred copy number, the samples in our study are comprised of highly varied mixtures of normal and DCIS cells, indicating that analyses of these data without this step would be misleading. We found that most of the DCIS lesions in the current study are composed of a mixture of multiple genetic clones that exist in a range of differentiation states. We observed notable differences in both the expression and relative abundance of cell states and substates between normal and DCIS cells, as well as between ER+ and ER− DCIS cells. Looking at the intersection among cell states and CNV-based subclones, we observe that subclones are frequently comprised of multiple cell states. This would seem to indicate that many of the subclones maintain the ability to differentiate.

Cells in the basal/myoep spectrum are significantly less abundant in the neoplastic cell populations, indicating that cell differentiation into this cell state is uncommon in DCIS. However, ER− DCIS (but not ER+) retained a significant proportion of cells with basal characteristics (LumSec-basal substate) distinct from the mature myoepithelial population. We conclude that each DCIS lesion has variable compositions of cell substates that only in part mimic the distribution and heterogeneity of the normal breast, with ER− DCIS (compared with ER+ DCIS) more closely resembling the cell distribution in the normal breast. In a DCIS cohort with longitudinal outcome, we found that high levels of LumSec-major and low levels of LumSec-myo substates are associated with shorter time to IBC progression. LumSec-major cells express *MMP7,* which may promote invasiveness, whereas LumSec-myo cells share properties with mature myoepithelial cells and may delay progression (34).

Previous studies have found that attenuation of the basal/myoep cell layer is common in DCIS and has been considered as one of a number of potential mechanisms of invasion (13–15). However, whereas invasive cancers lose the basal/myoep layer, we recently showed, using spatial proteomics, that attenuation of the basal/myoep layer (defined by ACTA2 expression) is not associated with longitudinal progression of DCIS, suggesting that there may not be a direct cause and effect relationship (58). However, our current results show that the changes in cell composition related to the expansion of DCIS cells may affect critical epithelial function, in particular the structure and composition of the BM.

Some of the essential functions of the BM in maintaining integrity include providing a scaffold for epithelial cells, helping them maintain their shape and resist mechanical stress, and anchoring the epithelial cells to the underlying matrix (50, 59, 60). In a mouse DCIS model, BM maintenance was shown to be an ongoing and active process (49). Our scRNA-seq data support this as well. These findings are consistent with previous studies on synchronous DCIS and IBC lesions that found epithelial gene expression changes indicative of progressive loss of basal layer integrity (61). Using a 3D *in vitro* culture model that recapitulates the structure and BM of the breast duct, we demonstrated that compromising BM integrity by degrading COL4 in these *in vitro* acini increased the invasive phenotype. COL4 is one of the main structural proteins in the BM, and its importance in maintaining epithelial integrity is illustrated by its evolutionary appearance at the transition from unicellular to multicellular organisms (62). Notably in our investigation of BM continuity in DCIS adjacent to areas of microinvasion, we did not see COL4 loss in DCIS, rather it was lost only in areas in which there was histologic evidence of invasion. These results are consistent with COL4 being an essential component of the BM, the loss of its integrity being tightly associated with the invasive phenotype.

In summary, our single-cell data, supplemented by evidence from an *in vitro* model, support that overgrowth of neoplastic cell populations and the relative reduction of BM-producing cells within the epithelium may create an unstable epithelial structure. This challenges the prevailing dogma, reframing the process of invasion as a loss-of-function event due to an imbalance in critical cell populations essential for BM integrity during the neoplastic process, rather than a gain of function by the neoplastic cells. This model helps to explain the previously described genomic identity between epithelial cells in DCIS and invasive cancer and suggests novel alternative targets for future interception efforts that focus on maintenance of the BM as an adjunct to targeting of the epithelial compartment.

Authors' Disclosures

X. Qin reports grants from U2C CA-17-035 during the conduct of the study. M.R. Lee reports grants from U2C CA-17-035 during the conduct of the study. A. Hall reports grants from the NIH during the conduct of the study. G.A. Colditz reports grants from the Breast Cancer Research Foundation during the conduct of the study. E.S. Hwang reports grants from the NIH during the conduct of the study. J.R. Marks reports grants from the NIH during the conduct of the study. K. Owzar reports grants from U2C CA-17-035 during the conduct of the study. No disclosures were reported by the other authors.

Disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Authors' Contributions

X. Qin: Data curation, software, formal analysis, validation, investigation, visualization, methodology, writing–original draft, writing–review and editing. **S.H. Strand:** Resources, data curation, software, formal analysis, validation, investigation, visualization, methodology, writing–original draft, writing–review and editing. **M.R. Lee:** Data curation, software, formal analysis, validation, investigation, visualization, methodology, writing–original draft, writing–review and editing. **A. Saraswathibhatla:** Resources, formal analysis, investigation, visualization, methodology, writing–original draft,

writing–review and editing. **D.G. van IJzendoorn:** Formal analysis, investigation, writing–review and editing. **C. Zhu:** Data curation, investigation, methodology, writing–review and editing. **S. Vennam:** Resources, Data curation, methodology, writing–review and editing. **S. Varma:** Resources, methodology, writing–review and editing. **A. Hall:** Resources, supervision, writing–review and editing. **R.E. Factor:** Resources, formal analysis, supervision, visualization, methodology, writing–original draft, writing–review and editing. **L. King:** Resources, formal analysis, supervision, investigation, writing–original draft, writing–review and editing. **L. Simpson:** Conceptualization, resources, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, writing–original draft, project administration, writing–review and editing. **X. Luo:** Resources, formal analysis, investigation, visualization, methodology, writing–original draft, writing–review and editing. **G.A. Colditz:** Conceptualization, resources, formal analysis, supervision, funding acquisition, investigation, writing–original draft, project administration, writing–review and editing. **S. Jiang:** Resources, formal analysis, supervision, investigation, visualization, methodology, writing–original draft, writing–review and editing. **O. Chaudhuri:** Resources, formal analysis, supervision, investigation, writing–original draft, writing–review and editing. **E.S. Hwang:** Conceptualization, resources, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, writing–original draft, project administration, writing–review and editing. **J.R. Marks:** Conceptualization, resources, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, writing–original draft, project administration, writing–review and editing. **K. Owzar:** Conceptualization, resources, data curation, software, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, writing–original draft, project administration, writing–review and editing. **R.B. West:** Conceptualization, resources, data curation, software, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, writing–original draft, project administration, writing–review and editing.

Acknowledgments

Research reported in this publication was supported by the NCI of the NIH under award number U2C CA-17-035 Pre-Cancer Atlas Research Centers (E.S. Hwang, M.R. Lee, J.R. Marks, K. Owzar, S.H. Strand, R.B. West, and X. Qin) and CA014236 (M.R. Lee, K. Owzar, and X. Qin). E.S. Hwang and G.A. Colditz received support from the Breast Cancer Research Foundation.

Note

Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Received August 26, 2024; revised January 16, 2025; accepted March 12, 2025; posted first March 18, 2025.

References

1. Virnig BA, Tuttle TM, Shamliyan T, Kane RL. Ductal carcinoma in situ of the breast: a systematic review of incidence, treatment, and outcomes. J Natl Cancer Inst 2010;102:170–8.
2. Mannu GS, Wang Z, Broggio J, Charman J, Cheung S, Kearins O, et al. Invasive breast cancer and breast cancer mortality after ductal carcinoma in situ in women attending for breast screening in England, 1988–2014: population based observational cohort study. BMJ 2020;369:m1570.
3. Biermann J, Parris TZ, Nemes S, Danielsson A, Engqvist H, Werner Rönnerman E, et al. Clonal relatedness in tumour pairs of breast cancer patients. Breast Cancer Res 2018;20:96.
4. Lips EH, Kumar T, Megalios A, Visser LL, Sheinman M, Fortunato A, et al. Genomic analysis defines clonal relationships of ductal carcinoma in situ and recurrent invasive breast cancer. Nat Genet 2022;54:850–60.
5. Waldman FM, DeVries S, Chew KL, Moore DH II, Kerlikowske K, Ljung BM. Chromosomal alterations in ductal carcinomas in situ and their in situ recurrences. J Natl Cancer Inst 2000;92:313–20.
6. Allred DC, Wu Y, Mao S, Nagtegaal ID, Lee S, Perou CM, et al. Ductal carcinoma in situ and the emergence of diversity during breast cancer evolution. Clin Cancer Res 2008;14:370–8.
7. Strand SH, Rivero-Gutiérrez B, Houlahan KE, Seoane JA, King LM, Risom T, et al. Molecular classification and biomarkers of clinical outcome in breast ductal carcinoma in situ: analysis of TBCRC 038 and RAHBT cohorts. Cancer Cell 2023;41:1381.
8. Shah C, Bremer T, Cox C, Whitworth P, Patel R, Patel A, et al. The clinical utility of DCISionRT® on radiation therapy decision making in patients with ductal carcinoma in situ following breast-conserving surgery. Ann Surg Oncol 2021;28:5974–84.
9. Solin LJ, Gray R, Baehner FL, Butler SM, Hughes LL, Yoshizawa C, et al. A multigene expression assay to predict local recurrence risk for ductal carcinoma in situ of the breast. J Natl Cancer Inst 2013;105:701–10.
10. Casasent AK, Schalck A, Gao R, Sei E, Long A, Pangburn W, et al. Multiclonal invasion in breast tumors identified by topographic single cell sequencing. Cell 2018;172:205–17.e12.
11. Porter DA, Krop IE, Nasser S, Sgroi D, Kaelin CM, Marks JR, et al. A SAGE (serial analysis of gene expression) view of breast tumor progression. Cancer Res 2001;61:5697–702.
12. Ma X-J, Salunga R, Tuggle JT, Gaudet J, Enright E, McQuary P, et al. Gene expression profiles of human breast cancer progression. Proc Natl Acad Sci U S A 2003;100:5974–9.
13. Burstein HJ, Polyak K, Wong JS, Lester SC, Kaelin CM. Ductal carcinoma in situ of the breast. N Engl J Med 2004;350:1430–41.
14. Adriance MC, Inman JL, Petersen OW, Bissell MJ. Myoepithelial cells: good fences make good neighbors. Breast Cancer Res 2005;7:190–7.
15. Chang J, Chaudhuri O. Beyond proteases: basement membrane mechanics and cancer invasion. J Cell Biol 2019;218:2456–69.
16. Tokura M, Nakayama J, Prieto-Vila M, Shiino S, Yoshida M, Yamamoto T, et al. Single-cell transcriptome profiling reveals intratumoral heterogeneity and molecular features of ductal carcinoma in situ. Cancer Res 2022;82:3236–48.
17. Foley JW, Zhu C, Jolivet P, Zhu SX, Lu P, Meaney MJ, et al. Gene expression profiling of single cells from archival tissue with laser-capture microdissection and Smart-3SEQ. Genome Res 2019;29:1816–25.
18. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: open source software for digital pathology image analysis. Sci Rep 2017;7:16878.
19. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491–8.
20. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 2013;43:11.10.1–33.
21. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303.
22. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol 2017;35:316–9.
23. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol 2020;38:276–8.
24. Scheinin I, Sie D, Bengtsson H, van de Wiel MA, Olshen AB, van Thuijl HF, et al. DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. Genome Res 2014;24:2022–32.
25. Poell JB, Mendeville M, Sie D, Brink A, Brakenhoff RH, Ylstra B. ACE: absolute copy number estimation from low-coverage whole-genome sequencing data. Bioinformatics 2019;35:2847–9.
26. van de Wiel MA, Kim KI, Vosse SJ, van Wieringen WN, Wilting SM, Ylstra B. CGHcall: calling aberrations for array CGH tumor profiles. Bioinformatics 2007;23:892–4.
27. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21.
28. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;30:923–30.
29. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, et al. GENCODE 2021. Nucleic Acids Res 2021;49:D916–23.
30. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017;8:14049.

31. Hao Y, Hao S, Andersen-Nissen E, Mauck WM III, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184:3573–87.e29.
32. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM III, et al. Comprehensive integration of single-cell data. Cell 2019;177:1888–902.e21.
33. Guo H, Li J. scSorter: assigning cells to known cell types according to marker genes. Genome Biol 2021;22:69.
34. Kumar T, Nee K, Wei R, He S, Nguyen QH, Bai S, et al. A spatially resolved single-cell genomic atlas of the adult human breast. Nature 2023;620:181–91.
35. Paul I, Bolzan D, Youssef A, Gagnon KA, Hook H, Karemore G, et al. Parallelized multidimensional analytic framework applied to mammary epithelial cells uncovers regulatory principles in EMT. Nat Commun 2023;14:688.
36. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 2019;9:5233.
37. Paradis E, Schliep K. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 2019;35:526–8.
38. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.
39. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 2015;1:417–25.
40. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods 2017;14:979–82.
41. Qiu X, Hill A, Packer J, Lin D, Ma Y-A, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods 2017;14:309–15.
42. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014;32:381–6.
43. Rodriguez A, Laio A. Machine learning. Clustering by fast search and find of density peaks. Science 2014;344:1492–6.
44. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.
45. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol 2019;37:773–82.
46. Therneau TM, Li H. Computing the Cox model for case cohort designs. Lifetime Data Anal 1999;5:99–112.
47. Chang J, Saraswathibhatla A, Song Z, Varma S, Sanchez C, Alyafei NHK, et al. Cell volume expansion and local contractility drive collective invasion of the basement membrane in breast cancer. Nat Mater 2024;23:711–22.
48. Debnath J, Muthuswamy SK, Brugge JS. Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. Methods 2003;30:256–68.
49. Morgner J, Bornes L, Hahn K, López-Iglesias C, Kroese L, Pritchard CEJ, et al. A Lamb1Dendra2 mouse model identifies basement-membrane-producing origins and dynamics in PyMT breast tumors. Dev Cell 2023;58:535–49.e5.
50. Pozzi A, Yurchenco PD, Iozzo RV. The nature and biology of basement membranes. Matrix Biol 2017;57–58:1–11.
51. Pöschl E, Schlötzer-Schrehardt U, Brachvogel B, Saito K, Ninomiya Y, Mayer U. Collagen IV is essential for basement membrane stability but dispensable for initiation of its assembly during early development. Development 2004;131:1619–28.
52. Ghannam SF, Rutland CS, Allegrucci C, Mongan NP, Rakha E. Defining invasion in breast cancer: the role of basement membrane. J Clin Pathol 2023;76:11–8.
53. Sternlicht MD, Lochter A, Sympson CJ, Huey B, Rougier JP, Gray JW, et al. The stromal proteinase MMP3/stromelysin-1 promotes mammary carcinogenesis. Cell 1999;98:137–46.
54. Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, et al; Gene Ontology Consortium. The gene ontology knowledgebase in 2023. Genetics 2023;224:iyad031.
55. Lee JY, Chang JK, Dominguez AA, Lee H-P, Nam S, Chang J, et al. YAP-independent mechanotransduction drives breast cancer progression. Nat Commun 2019;10:1848.
56. Pant SM, Belitskin D, Ala-Hongisto H, Klefström J, Tervonen TA. Analyzing the type II transmembrane serine protease hepsin-dependent basement membrane remodeling in 3D cell culture. Methods Mol Biol 2018;1731:169–78.
57. Visvader JE, Stingl J. Mammary stem cells and the differentiation hierarchy: current status and perspectives. Genes Dev 2014;28:1143–58.
58. Risom T, Glass DR, Averbukh I, Liu CC, Baranski A, Kagel A, et al. Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. Cell 2022;185:299–310.e18.
59. Yurchenco PD. Basement membranes: cell scaffoldings and signaling platforms. Cold Spring Harbor Perspect Biol 2011;3:a004911.
60. Li S, Harrison D, Carbonetto S, Fassler R, Smyth N, Edgar D, et al. Matrix assembly, regulation, and survival functions of laminin and its receptors in embryonic stem cell differentiation. J Cell Biol 2002;157:1279–90.
61. Rebbeck CA, Xian J, Bornelöv S, Geradts J, Hobeika A, Geiger H, et al. Gene expression signatures of individual ductal carcinoma in situ lesions identify processes and biomarkers associated with progression towards invasive ductal carcinoma. Nat Commun 2022;13:3399.
62. Fidler AL, Darris CE, Chetyrkin SV, Pedchenko VK, Boudko SP, Brown KL, et al. Collagen IV and basement membrane at the evolutionary dawn of metazoan tissues. Elife 2017;6:e24176.