Single Cell Expression Analysis of Ductal Carcinoma in Situ Identifies Alterations in Epithelial Integrity

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

[1] Duke Cancer Institute, Duke University School of Medicine, Durham, NC 27705, USA
[2] Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA
[3]Department of Pathology, Duke University School of Medicine, Durham, NC 27708, USA
[4] Department of Surgery, Duke University School of Medicine, Durham, NC 27708, USA
[5] Department of Surgery, Washington University School of Medicine, St. Louis, MO 63110, USA

[6] Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27705, USA

#These authors contributed equally
†These authors contributed equally
*Correspondence: jeffrey.marks@duke.edu (J.R.M), kouros.owzar@duke.edu (K.O.), rbwest@stanford.edu (R.B.W.)

ABSTRACT

The defining characteristic of ductal carcinoma in situ (DCIS) is neoplastic epithelial growth within a duct surrounded by an intact basement membrane (BM) whereas invasive breast cancer (IBC) is defined by tumor cell extravasation associated with loss of the intact BM. To understand mechanisms underlying the development of DCIS, we performed single-cell RNA-sequencing (scRNA-seq) on DCIS and synchronous normal tissue. We identified neoplastic epithelial cells by inferring copy number variations from RNA expression and then classified cells using previously defined phenotypes from normal human breast epithelium. We identify an increase in luminal cells in DCIS and a concomitant decrease in basal cells, which are a significant source of BM gene expression. We hypothesize that the reduction of the relative abundance of basal cells reduces the integrity of the BM that can result in the invasive phenotype.

Ductal carcinoma in situ (DCIS) of the breast is a non-invasive condition commonly identified through mammographic screening. A primary diagnosis of DCIS carries little mortality risk by itself but its presence is a risk factor for subsequent invasive cancer [1,2]. When invasive breast cancer (IBC) does arise in the same breast, there is a clear clonal genetic relationship to the primary DCIS in approximately 80% of cases indicative of clonal progression rather than a de novo process [3-5].

Pathology characterizes DCIS as a neoplastic proliferation of epithelial cells within an intact basement membrane. As a group, DCIS often share the same constellation of genetic abnormalities found in IBC [6]. Moreover, many IBCs have co-existing DCIS components and these synchronous lesions are usually genetically related [7-9].

Mechanisms underlying the progression from DCIS to IBC are not well understood although single cell genetic evidence indicates that polyclonal invasion is commonly observed [10]. Numerous studies of normal or benign human breast tissue indicate that ducts and lobules are composed of different subtypes of epithelial cells [11,12]. The well-established intrinsic phenotypes of IBC appear to arise from these various normal epithelial cell subtypes based on similarities of expressed genes [13]. We and others have studied expression patterns of DCIS to determine whether the IBC subtypes can be applied to this pre-invasive state. The PAM50 classifier identifies each of the five major subtypes (luminal A, luminal B, basal, HER2, and normal-like) however the triple negative basal subtype is not clearly represented in DCIS [14,15].

Single-cell RNA sequencing (scRNA-seq) has been established as a valuable tool to understand tissue composition at single-cell and whole-transcriptome resolution. A recent application of this approach to DCIS concluded that significant intra-tumor heterogeneity exists in the epithelial cell populations [16]. One of the primary limitations of scRNA-seq is that spatial information is lost which is a major concern with the typical surgical specimens that contain an admixture of DCIS and normal breast epithelia.

As part of the Human Tumor Atlas Network (HTAN), we focused on changes in cellular composition and phenotypes that occur between patient-matched normal breast and DCIS starting from scRNA-seq data generated on DCIS and normal breast tissue. We observed a consistent loss of basal cells in the single cell DCIS specimens and extended the study to bulk RNA-seq data sets and specific in situ analyses of progressed and non-progressed lesions. We hypothesize that progressive loss of basal/myoepithelial cells results in reduced production of specific components of the BM, leading to subsequent invasion.

RESULTS

Identification of neoplastic cells from scRNA-seq data

We performed scRNA-seq on DCIS (n=18) and matched synchronous normal tissues (n=12) from 15 patients undergoing mastectomy (**Extended data Table 1**). Both the DCIS and normal specimens were bisected, with one half used for single-cell dissociation, and the other half formalin fixed and embedded in paraffin to confirm the presence and extent of DCIS and normal breast epithelium (specimen and data generation workflow shown in **Figure 1A**). All specimens containing DCIS also contained an admixture of normal breast ductal structures. From the FFPE samples we performed whole genome sequencing (WGS) and "bulk" RNA sequencing on microdissected areas of DCIS. A schematic of the bioinformatics workflow to classify cells based on normal breast cell types and identify DCIS versus Normal epithelial cells is shown in **Figure 1B**.

Two-dimensional UMAP embedding plots of cells from DCIS and normal tissue samples colored by cluster membership are shown in **Figure 2A**. Comparison of the cell clusters identified in the normal samples and the DCIS samples showed that the overall architecture of the clustering was preserved between the two sample types, but some clusters were only present in DCIS samples from specific patients (**Extended data Figure 1**).

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 non-neoplastic epithelial cells. To identify whether epithelial cells were from regions of DCIS or normal tissue, we used the scRNA expression to infer cell-specific copy number variation (CNV) profiles and compared those to copy number profiles derived from microdissected DCIS from the adjacent paraffin block (**Figure 2B**). This workflow enabled us to identify three groups of cells, termed "DD" (DCIS cells from DCIS samples), "DN" (Normal cells from DCIS samples), and "NN" (Normal cells from Normal breast samples).

To demonstrate that our CNV-based approach distinguished DD from DN epithelial cells we compared aggregated gene expression (pseudo-bulk samples) between NN and DD cell populations. A heat map of the top 250 significantly up-regulated and 250 down-regulated genes comparing the DCIS (DD) to normal cells (NN) is shown in **Figure 2C**. DN cells were not used in the differential expression (DEseq) analysis but cluster closely with cells from normal breast (NN) indicating that our expression based copy number system is performing a functional separation. DEseq comparison of DN vs NN analysis yielded just 9 significant genes (adj. $p < 0.05$). As a further test, we performed pathway analysis on genes differentially expressed between all three groups, DD, DN, and NN (**Extended data Figure 2**). The same pathways were found to be significant comparing DD to both NN and DN cells including cell cycle progression and metabolic pathways elevated in DD further supporting the accuracy of the cell level categorization as neoplastic and non-neoplastic.

Comparison of epithelial sub-type proportions by clinical ER status

Kumar et al. recently identified 3 main epithelial cell states and 11 substates using scRNAseq analysis of normal breast tissue [12]. From among 146,146 cells which passed our final quality control metrics, we classified 45,908 cells from DCIS and Normal libraries into three inferred primary epithelial states (Basal, Luminal secretory (LumSec), and Luminal hormone responsive (LumHR)) and 11 inferred substates. The 11 substates map to the three primary states in normal breast epithelial cells, providing higher resolution cell type annotations. In the previously published normal breast atlas data, the Basal substate is equivalent to the primary Basal state whereas LumSec is comprised of seven substates and LumHR is comprised of three substates. In our scRNA dataset, the three LumHR substates had similar gene expression profiles while the seven LumSec substates were more diverse (**Extended data Figure 3A).** The three major states in the context of our DCIS and normal were displayed in epithelial-specific UMAPs (**Figure 3A**). We observed an overall loss of LumSec and Basal cells within the population of DCIS cells (DD) compared to the normal population (DNNN). At the patient level, there was a reduction in both the proportion and the total number of Basal cells in both ER+ and ER- DCIS compared to normal samples (**Figure 3B**, adj. p = 0.009, 0.027, respectively). LumHR cells comprised most of the transformed epithelial cells from ER+ DCIS (adj. p = 0.045, 0.045) and, concomitantly, there were very few LumSec and Basal cells in these samples. The DD ER- cells exhibited higher LumSec fractions compared to the DD ER+ (adj. p = 0.045). Aggregating the cell data showed highly significant distributional differences as well (**Figure 3C**). Matched by patient, ER- DCIS (compared to ER+) more closely resembled the distribution that was characteristic of the non-neoplastic breast cells (**Extended data Figure 3B**).

Applying the 11 epithelial substates to our data highlighted specific differences between DCIS (ER+ vs. ER-) and normal breast (**Figure 3D-F**). With respect to the LumSec cells, ER- DCIS retained a substantial proportion of the Basal substate and had elevated levels of the LumSec-prolif substate compared to normal cells (p = 0.007 **Figure 3E, Figure 3F**), accounting for the overall higher levels of LumSec cells in ER- compared to ER+ DCIS. The LumHR-active substate was significantly elevated in ER+ compared to both normal and ER- DCIS. ER- DCIS contained substantial proportions of the various luminal hormone responsive subtypes (i.e. LumHR-major and LumHR-SCGB) while ER+ DCIS contained very few cells classified as any LumSec substate. Therefore, even though the ER- DCIS cells do not express hormone receptor, a proportion of the transformed cells retained expression profiles that classified them as LumHR.

Trajectory analysis of epithelial subtypes

We explored the progression of the differentiated states of the epithelial cells using a trajectory analysis (**Figure 3G**). The full branching pattern was established using all epithelial cells from DCIS and normal samples and then divided into the three categories, DNNN, DD ER- and DD ER+. Predicted pseudotime is displayed on the top three panels and the cell identities are colored on the lower panels for both the three major epithelial cell states (middle panels) and the 11 substates (bottom panels). Basal cells occupied a branch at the earliest pseudotime and luminal cells (LumSec and LumHR) were present

at the other end of this scale. For the three major epithelial categories, there were fewer cells in the early cell states in DD ER- as compared to DNNN. In contrast, the ER+ cell states were more evenly distributed across the entire timeline. However, there was a relative increase in Basal cells at the beginning of the timeline in the DD ER- as compared to the DD ER+. The trajectories of the 11 epithelial cell substates showed similar distributions and demonstrated the relative positions of the substates along the timeline.

Validation and extension of findings in bulk-RNA libraries from independent DCIS cohorts

To extend and validate the findings from our scRNA-seq data, we used deconvolution to estimate the composition of epithelial cell states from bulk-RNA sequenced specimens including a series of matched synchronous samples (normal, DCIS, and IBC from the same patient (sample set 2)) and large longitudinal cohorts with known disease outcomes (sample set 3). We first confirmed the validity of this approach using the matched bulk-RNA-seq data from the FFPE DCIS specimens (**Figure 1A**). We found significantly lower relative proportion of Basal cells in DCIS stratified by ER status 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) (**Figure 4A)** replicating the specific findings in the matched scRNA-seq data (**Figure 3B**). At the individual sample level, we observed good correlation between scRNA-seq and FFPE bulk-seq data for all three epithelial cell types (**Extended data Figure 4A**).

The cohort of matched normal, DCIS, and IBC epithelial (LCM derived) samples (sample set 2) showed similar results. The relative proportions of Basal cells were significantly lower in DCIS vs normal breast for both ER+ and ER- DCIS (adj. $p < 0.001$, **Figure 4B**). In addition, LumHR cells were significantly enriched in ER+ DCIS compared to 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 to matched normal breast (adj. p = 0.008) but increased in ER- DCIS (adj. p = 0.011). Comparing the matched and synchronous DCIS to IBC, we observed minor increases in LumSec in ER+ and LumHR in ER+ IBC.

To investigate the association between the relative abundance of the cell substates and disease outcome, we analyzed the imputed cell fractions in retrospective DCIS case – control cohorts (sample set 3 [15]) of patients with a primary diagnosis of DCIS who later either did or did not have IBC progression. We found that patients with high levels of LumSec-major cells (p = 0.00022), and low levels of LumSec-myo cells (p = 0.019) were associated with shorter time to IBC recurrence (**Figure 4C and D and Extended data Figure 4B**).

Examination of basement membrane gene expression profiles

The general loss of basal cells in DCIS compared to normal tissue led us to consider their role in the synthesis of the basement membrane (BM). We examined the expression of canonical BM genes (GO0005604) [17] within the epithelial cell states and fibroblasts of normal and DCIS scRNA-seq samples (**Figure 5A, Extended data Figure 5**). As

expected, fibroblasts demonstrated relatively high expression of these genes. Basal cells demonstrated expression of a large number of BM genes, suggesting that this epithelial cell type also contributes to the integrity of the BM. While there was overlap, some BM genes were uniquely expressed in each cell type, such as LAMC2 in Basal cells. Luminal cells also had some BM gene expression although generally more restricted and at lower levels. LumSec-KIT cells highly expressed a number of BM genes (e.g., COL4A5) which were not expressed in fibroblasts (**Extended data Figure 5**). *In situ* analysis of protein and RNA expression of selected BM genes provided additional evidence that Basal cells contribute to the formation and maintenance of the BM (**Figure 5B and C**). *COL4A1* RNA expression was predominantly found in fibroblasts adjacent to the epithelial compartment (**Figure 5B**) with some expression also observed in Basal cells. Moreover, substantial expression of *COL4A1* was also evident in a subset of the ACTA2-positive cells (i.e. Basal cells) at the edge of the epithelial compartment. *LAMC2 RNA* was expressed in Basal cells but not fibroblasts (**Figure 5C**).

We investigated whether basement membrane integrity is associated with breast cancer invasion by examining the continuity of specific basement membrane components in a series of DCIS adjacent to areas of microinvasion (=<1 mm invasive component, sample set 4). LAMC2 expression was greatly decreased and sometimes absent in some ducts adjacent to the invasion despite the presence of myoepithelial/basal cells as defined by p63 (TP63) expression ((**Figure 6A and B (**p63 staining not shown)). COL4A1 protein was found in a consistent pattern around the ducts regardless of distance to the invasive component. However, LAMC2 continuity demonstrated a significant difference between DCIS close to IBC versus DCIS distant (p=0.04) (**Figure 6C**). In contrast, normal breast demonstrated a continuous COL4A1 and LAMC2 layer surrounding all identified ducts and lobules (**Extended data Figure 6A and B**).

Distribution of epithelial subpopulations in prostate neoplasia

To further investigate the distribution of epithelial subpopulations during cancer progression in other organs, we inferred epithelial cell states [18] from bulk RNA-seq data (sample set 5) consisting of LCM epithelial samples of matched normal prostate (n=51), prostate intraepithelial neoplasia (PIN, n = 46), and invasive prostate cancer (PCa, n=59). We observed the same trend as in breast neoplasia, with progressive loss of basal cells from normal to PIN (P = 0.06) to PCa (PIN vs PCa, P < 0.001, Normal vs PCa, P <0.001, **Extended data Figure 7**). This indicates that the observed loss of basal cells could be a shared phenomenon in cancers derived from organs with bilayered epithelium, such as the breast and prostate.

A conceptual model for compromise of epithelial integrity

A conceptual model of cancer progression that incorporates the consequences of decreased basal cell proportions in DCIS is presented in **Figure 7**. We hypothesize that the relative expansion of luminal cells and the resultant relative reduction of basal cells leads to a decrease in basal cell function. Since the Basal cells contribute to the

basement membrane this results in a compromise of epithelial integrity and loss of contact with an intact basement membrane that is interpreted as invasion.

DISCUSSION

Normal mammary ducts are composed of a number of cell types that exist along a differentiation spectrum starting from pluripotent stem cells that are capable of reconstituting full branching structures in cleared rodent fat pads [19]. Most cancers of the breast are thought to be clonal or oligoclonal which might be expected to give rise to a more restricted differentiation program. Given that DCIS retains the overall ductal architecture of the normal breast, measuring the cellular composition of these non-invasive cancers may reveal an intermediate state with respect to differentiation. In the current study, we identified epithelial cell states according to the recently published, most extensive single cell characterization of normal human breast cells [12]. Based on this theoretical framework, we compared the cellular composition of DCIS to normal breast. While we did detect significant changes in cell populations, DCIS are not monolithic with respect to differentiation state with ER- DCIS (compared to ER+) more closely resembling the distribution in the normal breast. For all DCIS in this study, there was a significant decrease in both the number and proportion of Basal cells that express genes associated with mature myoepithelial cells. However, ER- DCIS (but not ER+) retained a significant proportion of cells with basal characteristics distinct from the mature myoepithelial population. We conclude that DCIS have variable compositions of cell types that mimic the distribution and heterogeneity of the normal breast.

The epithelial cell category labels from Kumar et al. only partially coincide with preceding nomenclature [11,12]. A source of some confusion is whether basal and myoepithelial cells refer to the same cell entity. There are several categories of cells that express some combination of mature myoepithelial cell markers (e.g., smooth muscle actin) and basal type cytokeratins (e.g., KRT5). We conclude that mature myoepithelial cells are a subset of a somewhat heterogeneous category of cells with basal characteristics that have been termed “basal” cells. In particular, two cell categories, the LumSec-basal and basal cells, are known to produce a substantial number of basement membrane proteins.

Mature myoepithelial cells are present and surround many DCIS containing ductal structures [20]. Attenuation of this cell layer is common in DCIS and has been considered as a potential mechanism of invasion. Based on inferred copy number, our current analysis indicates that the vast majority of these cells are not part of the neoplastic process. This supports the long held idea that DCIS grows within the established ductal tree and that myoepithelial cells are residual from the normal breast. Alternatively, growth of the DCIS may also include a mixture of non-transformed breast progenitor cells that can produce these supporting cells.

While the expansion of the neoplastic population in DCIS is expected, we suggest that the resulting loss of basal cells may be a key step in progression to invasion. We hypothesize that the reduction of the relative abundance of basal cells reduces their contribution to the basement membrane and results in a loss of epithelial integrity. Some of the essential functions of the basement membrane in maintaining integrity include

providing a scaffold for the epithelial cells, helping them maintain their shape and resist mechanical stress, anchoring the epithelial cells to the underlying matrix, ensuring the stability of the tissue, and establishing and maintaining apical-basal polarity [21-24]. In a mouse DCIS model, basement membrane maintenance was shown to be an ongoing and active process [25]. COL4 is one of the main structural proteins in the basement membrane and its importance in maintaining epithelial integrity is illustrated by its evolutionary appearance at the transition from unicellular to multicellular organisms [26]. Notably, we did not see COL4 loss in DCIS, only when there was invasion. Overgrowth of mature cell populations and the relative decrease of basement membrane producing cells within the epithelium may create an unstable epithelial structure. This reframes the concept of invasion as a loss of function event due to critical cell populations being proportionally reduced in number during the neoplastic process, rather than a gain of function by the neoplastic cells (**Figure 7**). Our parallel analysis of prostate neoplasia supports the possibility that reduced basal cell differentiation may be a common feature of carcinomas. The loss of basement membrane function could be considered as a reversal of the evolutionary process to multicellularity.

Acknowledgements

Research reported in this publication was supported by National Cancer Institute of the National Institutes of Health under award number U2C CA-17-035 Pre-Cancer Atlas (PCA) Research Centers (E.S.H., M.R.L, J.R.M, K.O.,S.H.S., R.B.W., X.Q.).

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

Competing Interests

RBW has consulted for IonPath Inc.

Additional information

Supplementary Information is available for this paper: Supplementary File 1.

Correspondence and requests for materials should be addressed to Dr. Robert B. West (rbwest@stanford.edu).

Data and code availability

All data will be made publicly available on the HTAN data portal (https://humantumoratlas.org/).

All scripts needed to reproduce the analysis will be made available online by the time of publication.

1. Virnig, B.A., Tuttle, T.M., Shamliyan, T. & Kane, R.L. Ductal carcinoma in situ of the breast: a systematic review of incidence, treatment, and outcomes. *J Natl Cancer Inst* **102**, 170-8 (2010).
2. Mannu, G.S. *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* **369**, m1570 (2020).
3. Biermann, J. *et al.* Clonal relatedness in tumour pairs of breast cancer patients. *Breast Cancer Res* **20**, 96 (2018).
4. Lips, E.H. *et al.* Genomic analysis defines clonal relationships of ductal carcinoma in situ and recurrent invasive breast cancer. *Nat Genet* **54**, 850-860 (2022).
5. Waldman, F.M. *et al.* Chromosomal alterations in ductal carcinomas in situ and their in situ recurrences. *J Natl Cancer Inst* **92**, 313-20 (2000).
6. Allred, D.C. *et al.* Ductal carcinoma in situ and the emergence of diversity during breast cancer evolution. *Clin Cancer Res* **14**, 370-8 (2008).
7. Hernandez, L. *et al.* Genomic and mutational profiling of ductal carcinomas in situ and matched adjacent invasive breast cancers reveals intra-tumour genetic heterogeneity and clonal selection. *J Pathol* **227**, 42-52 (2012).
8. Kroigard, A.B. *et al.* Clonal expansion and linear genome evolution through breast cancer progression from pre-invasive stages to asynchronous metastasis. *Oncotarget* **6**, 5634-49 (2015).
9. Heselmeyer-Haddad, K. *et al.* Single-cell genetic analysis of ductal carcinoma in situ and invasive breast cancer reveals enormous tumor heterogeneity yet conserved genomic imbalances and gain of MYC during progression. *Am J Pathol* **181**, 1807-22 (2012).
10. Casasent, A.K. *et al.* Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell* **172**, 205-217 e12 (2018).
11. Nguyen, Q.H. *et al.* Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nature communications* **9**, 1-12 (2018).
12. Kumar, T. *et al.* A spatially resolved single-cell genomic atlas of the adult human breast. *Nature* **620**, 181-191 (2023).
13. Lim, E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med* **15**, 907-13 (2009).
14. Bergholtz, H. *et al.* Contrasting DCIS and invasive breast cancer by subtype suggests basal-like DCIS as distinct lesions. *NPJ Breast Cancer* **6**, 26 (2020).
15. Strand, S.H. *et al.* Molecular classification and biomarkers of clinical outcome in breast ductal carcinoma in situ: Analysis of TBCRC 038 and RAHBT cohorts. *Cancer Cell* **40**, 1521-1536 e7 (2022).
16. Tokura, M. *et al.* Single-Cell Transcriptome Profiling Reveals Intratumoral Heterogeneity and Molecular Features of Ductal Carcinoma In Situ. *Cancer Res* **82**, 3236-3248 (2022).
17. Gene Ontology, C. *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**(2023).
18. Henry, G.H. *et al.* A Cellular Anatomy of the Normal Adult Human Prostate and Prostatic Urethra. *Cell Rep* **25**, 3530-3542 e5 (2018).
19. Visvader, J.E. & Stingl, J. Mammary stem cells and the differentiation hierarchy: current status and perspectives. *Genes Dev* **28**, 1143-58 (2014).
20. Risom, T. *et al.* Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *Cell* **185**, 299-310 e18 (2022).
21. Yurchenco, P.D. Basement membranes: cell scaffoldings and signaling platforms. *Cold Spring Harb Perspect Biol* **3**(2011).

22. Pozzi, A., Yurchenco, P.D. & Iozzo, R.V. The nature and biology of basement membranes. *Matrix Biol* **57-58**, 1-11 (2017).
23. Hohenester, E. & Yurchenco, P.D. Laminins in basement membrane assembly. *Cell Adh Migr* **7**, 56-63 (2013).
24. Li, S. *et al.* Matrix assembly, regulation, and survival functions of laminin and its receptors in embryonic stem cell differentiation. *J Cell Biol* **157**, 1279-90 (2002).
25. Morgner, J. *et al.* A Lamb1Dendra2 mouse model identifies basement-membrane-producing origins and dynamics in PyMT breast tumors. *Dev Cell* **58**, 535-549 e5 (2023).
26. Fidler, A.L. *et al.* Collagen IV and basement membrane at the evolutionary dawn of metazoan tissues. *Elife* **6**(2017).

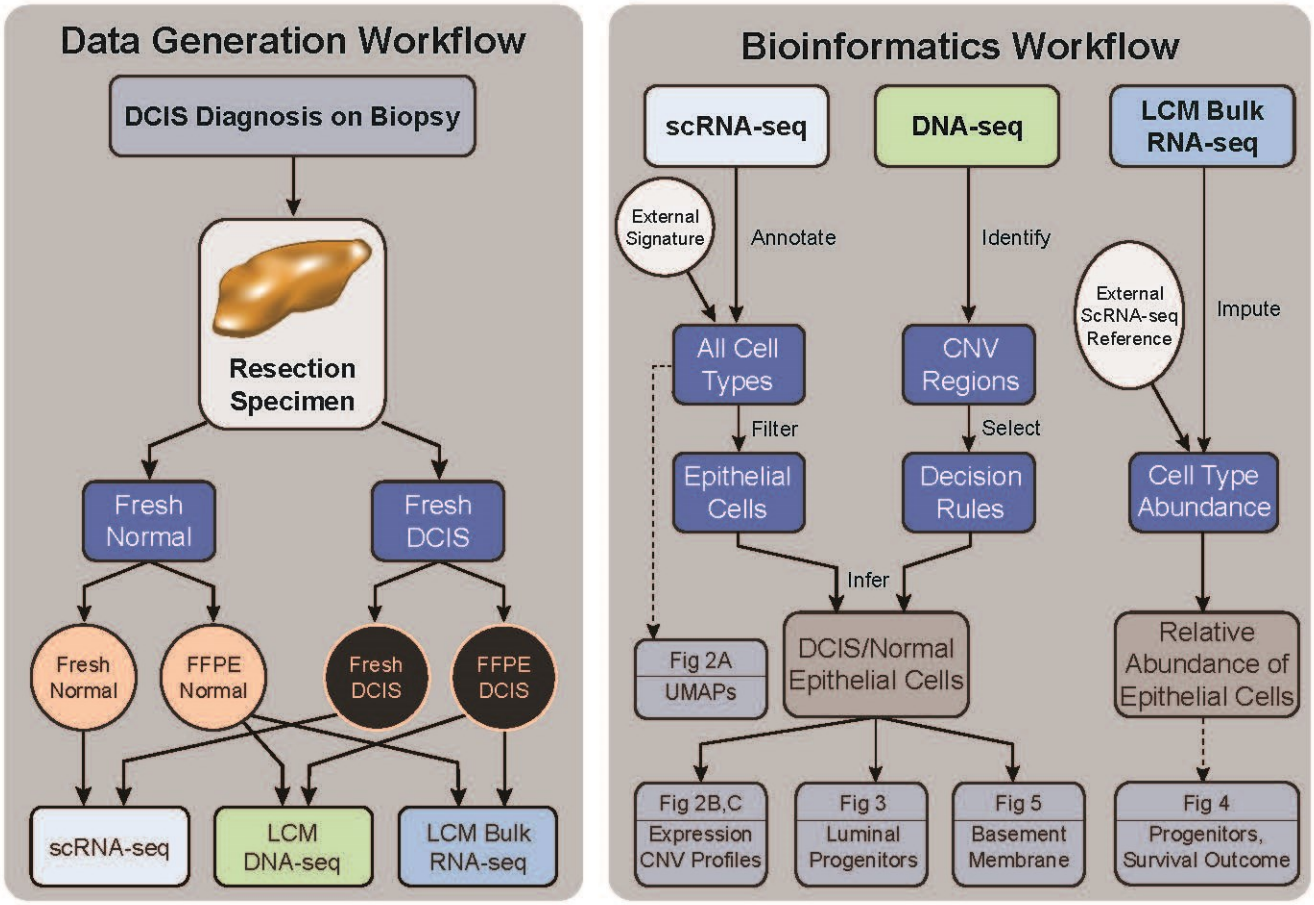


Figure 1. Overview of wet lab and informatic workflows.
A) Laboratory workflow to obtaining multi-modal sequencing data from resected specimens (Sample set 1 TA578 dataset) and **B)** bioinformatic workflow to infer cell phenotypes from scRNA and DNA data, infer cell fractions from specimen-matched bulk RNA data for downstream analyses.

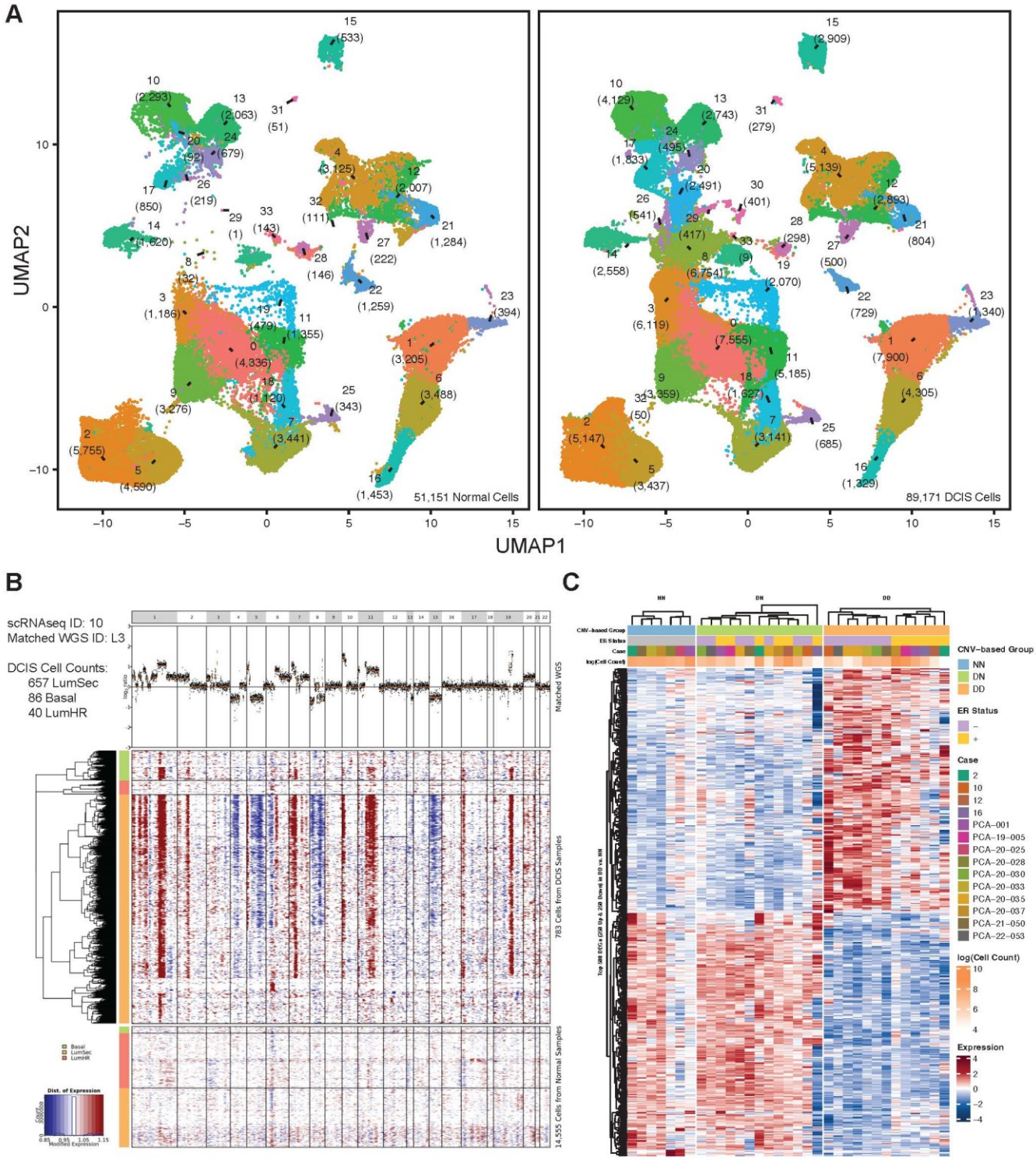


Figure 2. Identification and profiling of cells from scRNA-seq.
A) UMAP embedding plots based on single-cell transcriptome profiles from 140,322 cells across 13 DCIS sample libraries (89,171 cells) and 7 normal breast sample libraries (51,151 cells) obtained from 14 patients. Data integration and dimension reduction was performed on the entire collection of cells to identify 34 clusters. 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 cluster membership was used to color each coordinate. For each cluster, the ID is labeled at the median position and the number of cells is shown in the parentheses. **B)** A representative example of patient tumor-matched copy number motifs inferred from a low-pass WGS (upper subpanel) and inferred from single-cell DCIS (middle panel) libraries. A panel of 7 normal libraries were used as the reference (bottom panel). **C)** Relative expression of pseudo-bulk samples of cells from DCIS libraries inferred to be either DD (17,454 cells in 13 pseudo-bulk samples) or DN (13,899 cells from 13 pseudo-bulk samples), and cells from normal samples (NN, 14,555 cells in 7 pseudo-bulk samples). The top 250 up-regulated and 250 down-regulated differentially expressed genes (from DESeq2) ranked by adjusted *P*-value between DD and NN are displayed with expression of DN pseudo-bulk samples shown for comparison. Color bars indicate CNV-based group (DD, DN, NN), ER status (DD samples only), patient ID, and cell count (natural log scale).

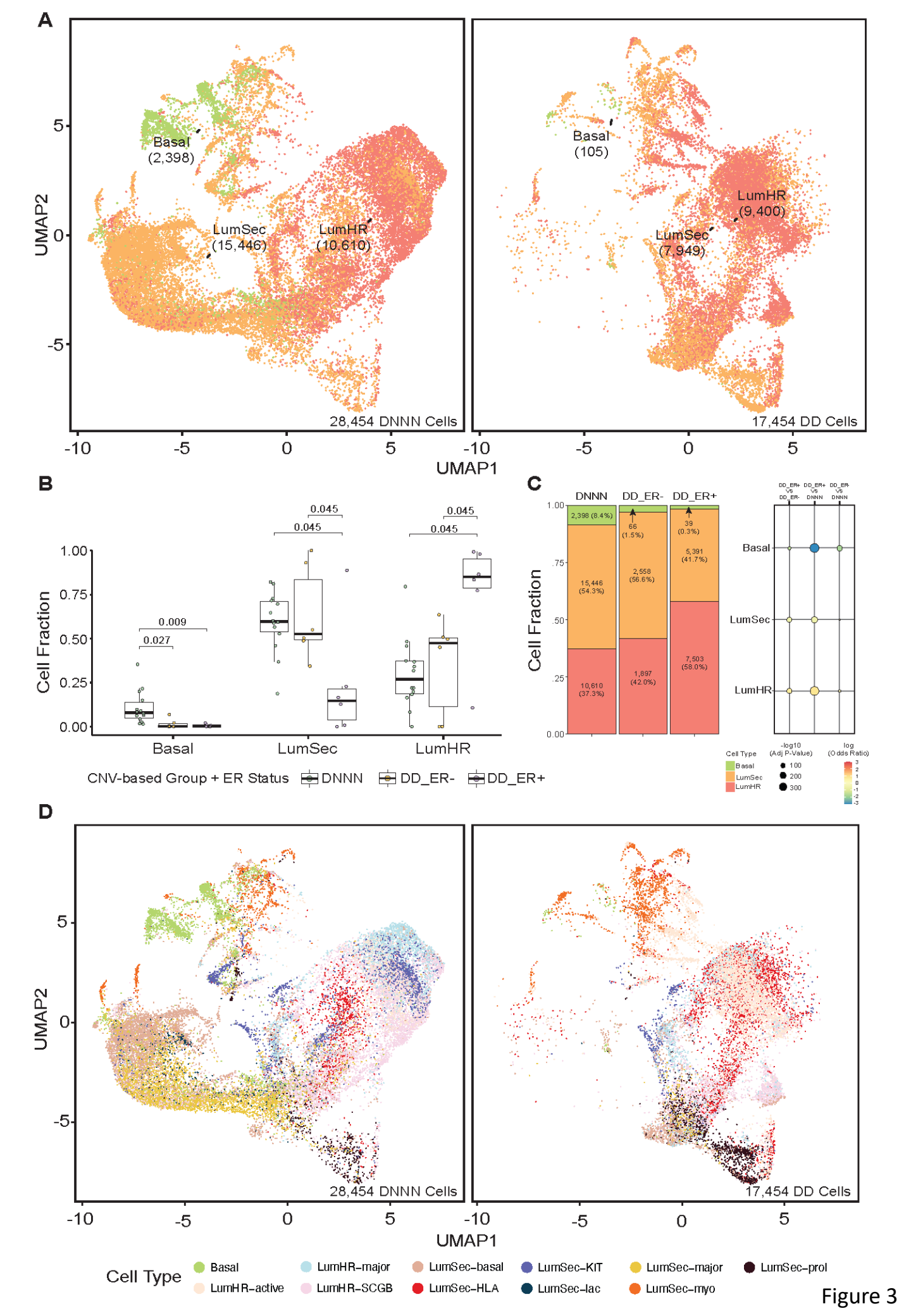
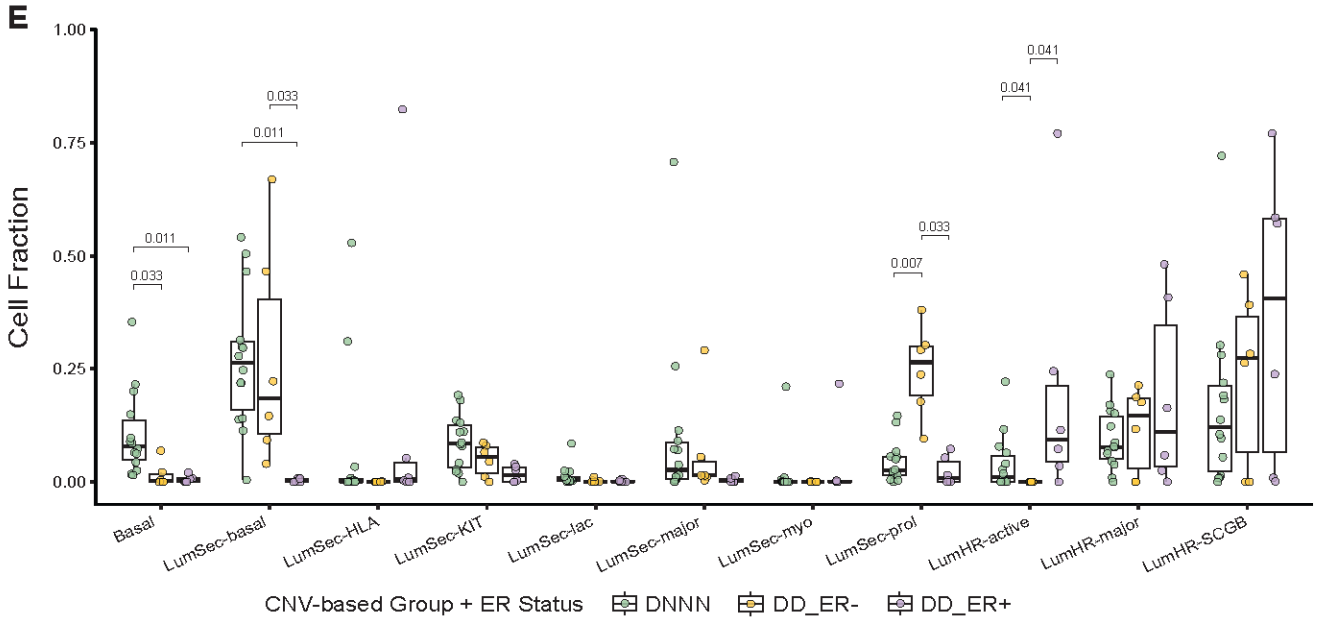
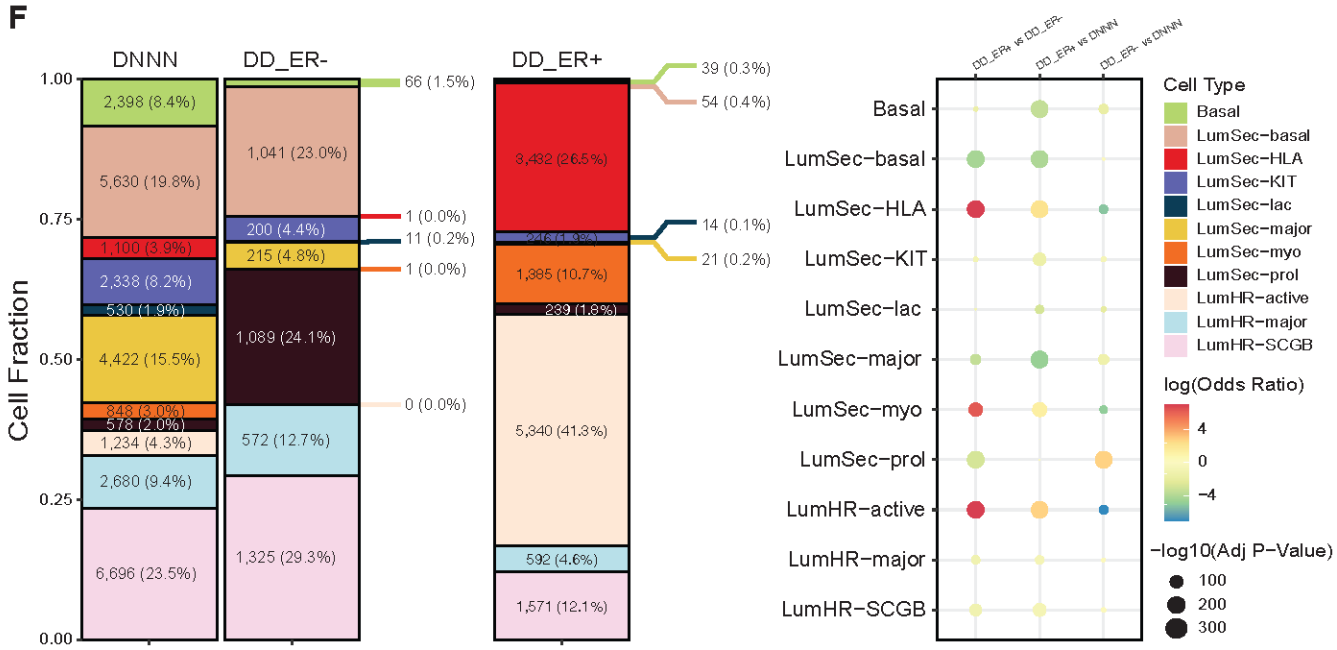

A
Basal
(2,398)
LumSec
(15,446)
LumHR
(10,610)
28,454 DNNN Cells
Basal
(105)
LumHR
(9,400)
LumSec
(7,949)
17,454 DD Cells
UMAP2
UMAP1
B
Cell Fraction
0.009
0.027
0.045
0.045
0.045
0.045
Basal
LumSec
LumHR
CNV-based Group + ER Status
DNNN
DD_ER-
DD_ER+
C
DNNN
DD_ER-
DD_ER+
2,398 (8.4%)
15,446 (54.3%)
10,610 (37.3%)
66 (1.5%)
2,558 (56.6%)
1,897 (42.0%)
39 (0.3%)
5,391 (41.7%)
7,503 (58.0%)
Cell Fraction
Basal
LumSec
LumHR
Cell Type
-log10 (Adj P-Value)
log (Odds Ratio)
D
28,454 DNNN Cells
17,454 DD Cells
UMAP2
UMAP1
Cell Type
Basal
LumHR-major
LumSec-basal
LumSec-KIT
LumSec-major
LumSec-prol
LumHR-active
LumHR-SCGB
LumSec-HLA
LumSec-lac
LumSec-myo

Figure 3


E
Cell Fraction
1.00
0.75
0.50
0.25
0.00
0.011
0.033
0.033
0.011
0.033
0.007
0.041
0.041
Basal
LumSec-basal
LumSec-HLA
LumSec-KIT
LumSec-lac
LumSec-major
LumSec-myo
LumSec-prol
LumHR-active
LumHR-major
LumHR-SCGB
CNV-based Group + ER Status
DNNN
DD_ER-
DD_ER+


F
DNNN
DD_ER-
DD_ER+
Cell Fraction
2,398 (8.4%)
5,630 (19.8%)
1,100 (3.9%)
2,338 (8.2%)
530 (1.9%)
4,422 (15.5%)
848 (3.0%)
578 (2.0%)
1,234 (4.3%)
2,680 (9.4%)
6,696 (23.5%)
1,041 (23.0%)
200 (4.4%)
215 (4.8%)
1,089 (24.1%)
572 (12.7%)
1,325 (29.3%)
66 (1.5%)
1 (0.0%)
11 (0.2%)
1 (0.0%)
0 (0.0%)
3,432 (26.5%)
1,385 (10.7%)
239 (1.8%)
5,340 (41.3%)
592 (4.6%)
1,571 (12.1%)
39 (0.3%)
54 (0.4%)
14 (0.1%)
21 (0.2%)
DD_ER+ vs DD_ER-
DD_ER+ vs DNNN
DD_ER- vs DNNN
Basal
LumSec−basal
LumSec−HLA
LumSec−KIT
LumSec−lac
LumSec−major
LumSec−myo
LumSec−prol
LumHR−active
LumHR−major
LumHR−SCGB
Cell Type
Basal
LumSec−basal
LumSec−HLA
LumSec−KIT
LumSec−lac
LumSec−major
LumSec−myo
LumSec−prol
LumHR−active
LumHR−major
LumHR−SCGB
log(Odds Ratio)
4
0
−4
−log10(Adj P−Value)
100
200
300

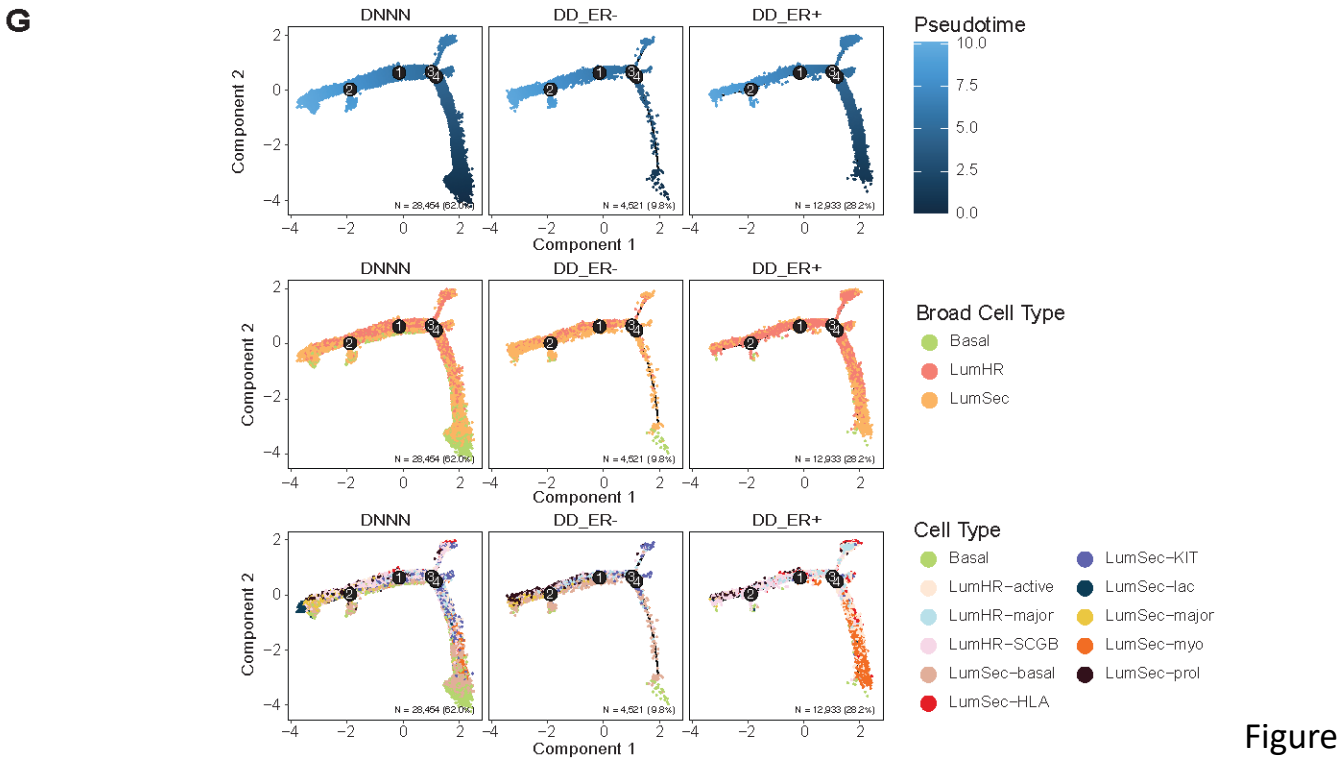
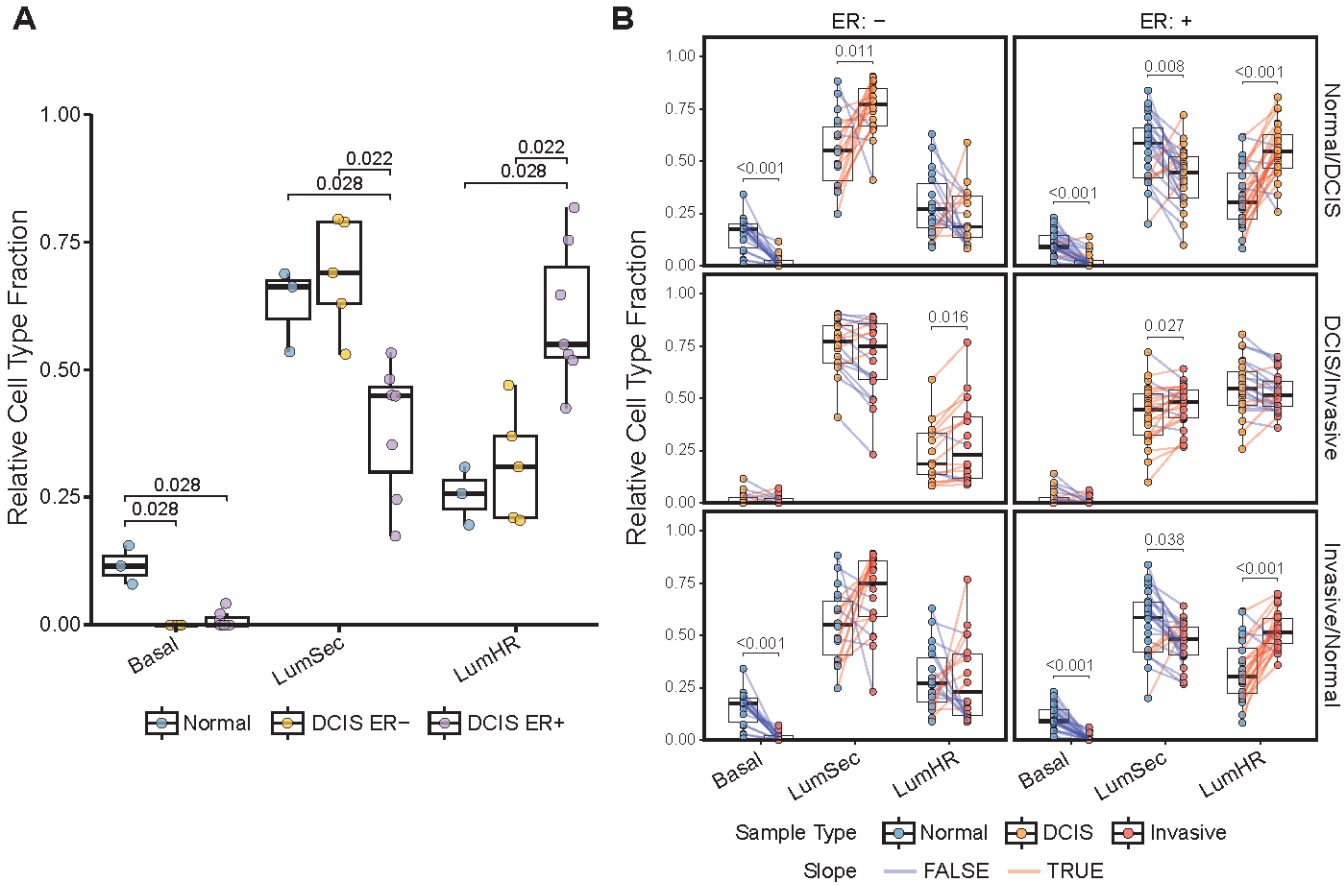

G
DNNN
DD_ER-
DD_ER+
Component 1
Component 2
Pseudotime
10.0
7.5
5.0
2.5
0.0
Broad Cell Type
Basal
LumHR
LumSec
Cell Type
Basal
LumHR−active
LumHR−major
LumHR−SCGB
LumSec−basal
LumSec−HLA
LumSec−KIT
LumSec−lac
LumSec−major
LumSec−myo
LumSec−prol

Figure 3

Figure 3. Comparison of epithelial subtype proportions.
A) UMAP embedding plots based on 45,908 epithelial cells (28,454 DNNN (left), 17,454 DD (right)) colored by the 3 major epithelial cell types found in normal breast [26]**.** Data integration and dimension reduction was performed on epithelial cells prior to UMAP embedding. **B)** Relative cell fractions of 3 major epithelial cell types at the pseudo-bulk level for DNNN, DD ER- and DD ER+. Variation in cell fraction by group was tested using Wilcoxon rank-sum and *P*-values were adjusted using Benjamini-Hochberg method to account for multiple comparisons. Results with an adjusted P-value of less than 0.05 are shown. **C)** Corresponding relative frequencies for the subtype compositions within DNNN, DD ER+, and DD ER- groups are summarized as stacked bars with text showing the number of cells and percentage in parentheses. The dot grid shows the resulting odds of subtype membership based on CNV and ER status. The color of the dots indicates the log odds ratio and the size indicates significance. Panels **D, E** and **F** present results at higher resolution, where 11 epithelial cell states are considered instead of the 3 major types. They correspond to panels **A**, **B**, and **C**, respectively. **G)** Cell trajectory analysis based on DNNN, DD ER- and DD ER+. Cells in the top row are colored by pseudotime, in the middle row cells they are colored by the 3 major epithelial cell types, and in the bottom row cells are colored by the 11 epithelial cell states.


A
Relative Cell Type Fraction
1.00
0.75
0.50
0.25
0.00
0.028
0.028
0.022
0.028
0.022
0.028
Basal
LumSec
LumHR
Normal
DCIS ER−
DCIS ER+
B
ER: −
ER: +
Normal/DCIS
DCIS/Invasive
Invasive/Normal
<0.001
0.011
0.008
<0.001
<0.001
0.016
0.027
<0.001
0.038
<0.001
<0.001
Basal
LumSec
LumHR
Sample Type
Normal
DCIS
Invasive
Slope
FALSE
TRUE

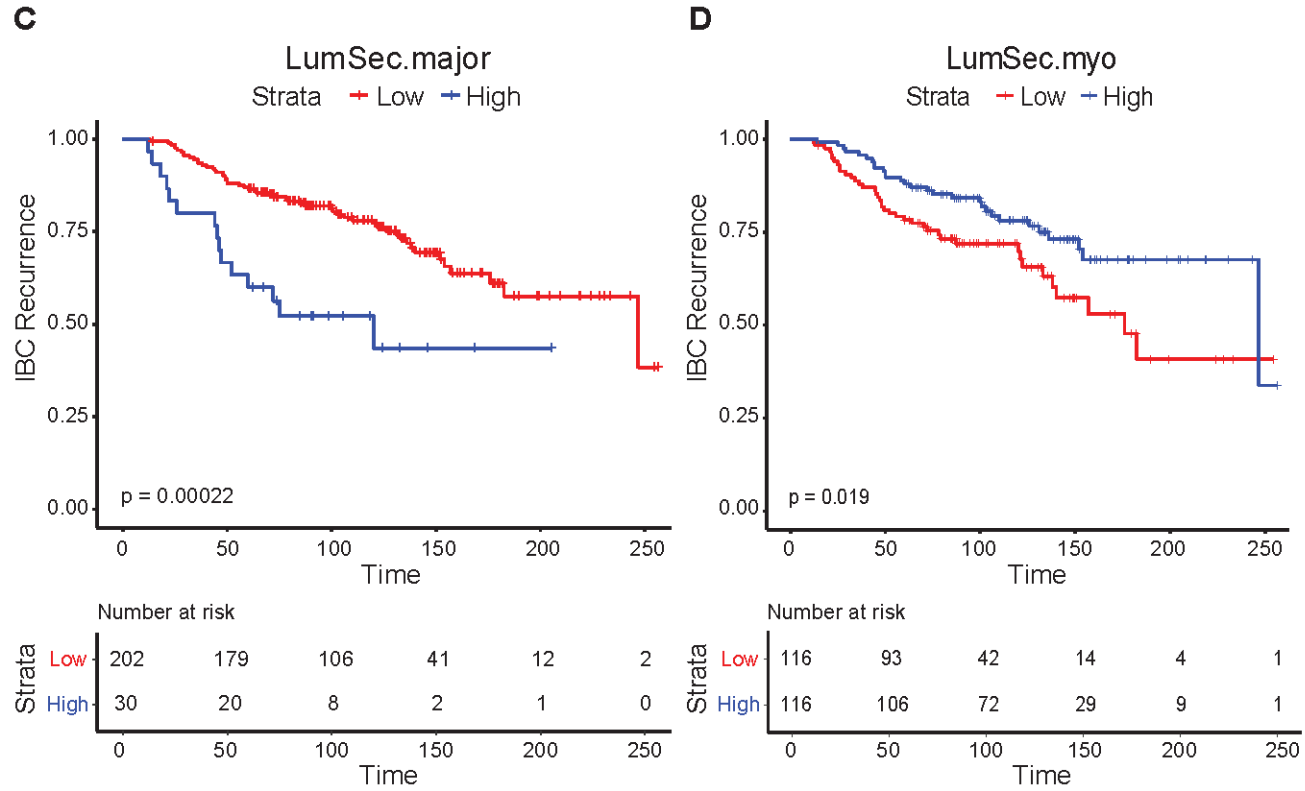
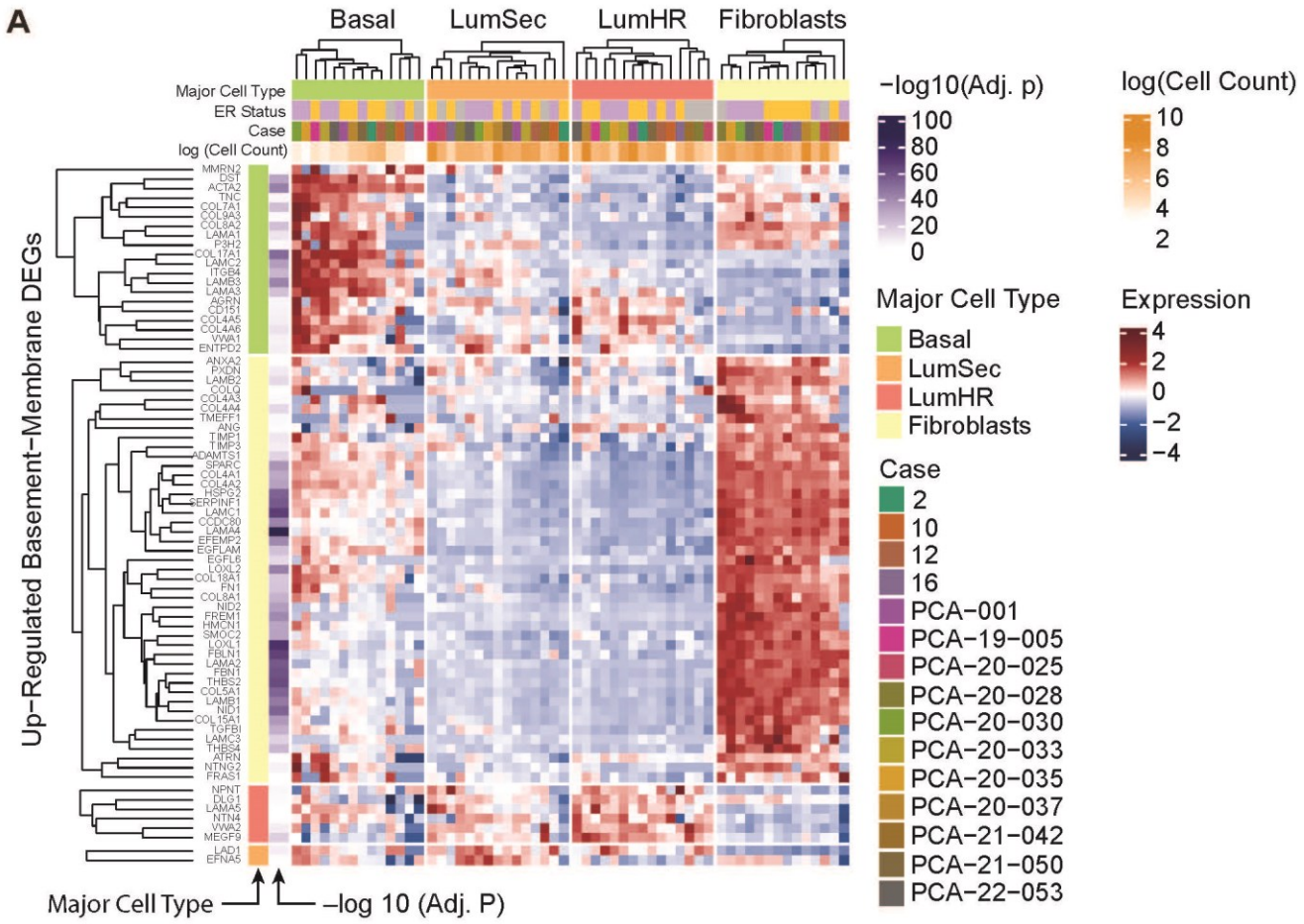
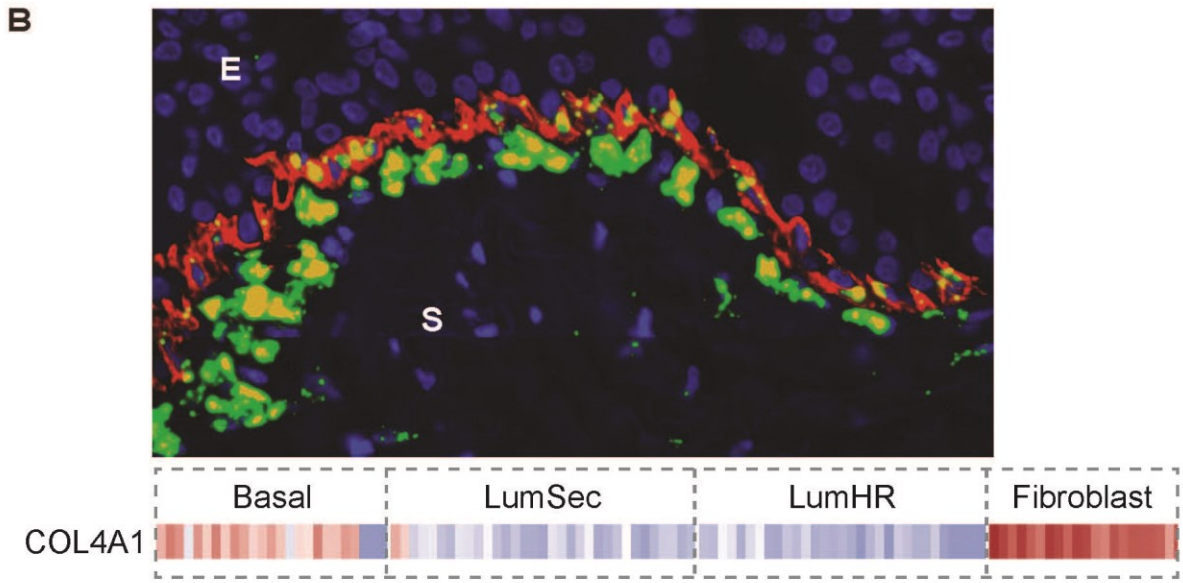

C
LumSec.major
Strata
Low
High
IBC Recurrence
1.00
0.75
0.50
0.25
0.00
p = 0.00022
0
50
100
150
200
250
Time
Number at risk
Strata
Low 202 179 106 41 12 2
High 30 20 8 2 1 0
D
LumSec.myo
Strata
Low
High
IBC Recurrence
p = 0.019
Number at risk
Low 116 93 42 14 4 1
High 116 106 72 29 9 1

Figure 4

Figure 4. Analysis of cell type proportions in independent DCIS cohorts.
Analysis of imputed cell fractions from two bulk-RNA sequencing datasets based on 3 major epithelial cell types found in normal breast [26]. The analyses were stratified by ER status for DCIS samples. **A)** Analysis of imputed cell fractions based on 3 major epithelial cell types found in normal breast [26] from bulk-RNA sequencing datasets of the FFPE sample set matching the scRNA-seq samples as depicted in Figure 1A with 3 Normal, 7 ER+ DCIS, and 5 ER- DCIS libraries obtained from 11 patients (Sample set 1 TA587 dataset). Differences were tested using the unpaired two-sample Wilcoxon rank-sum test. *P*-values were adjusted using Benjamini-Hochberg method to account for 9 comparisons. Results with an adjusted *P*-value of less than 0.05 are shown. **B)** Analysis of imputed cell fractions based on 3 major epithelial cell types found in normal breast [26] from bulk-RNA sequencing 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 S3S dataset). Observed increases (decreases) are colored in red (blue). Pairwise differences were tested using one-sample Wilcoxon signed-rank tests. *P*-values were adjusted using Benjamini-Hochberg method to account for 18 comparisons. **C, D)** Analysis of imputed cell fractions bulk-RNA sequencing datasets of retrospective DCIS case – control cohorts (Sample set 3 DCIS Atlas dataset [42]) demonstrating that high levels of LumSec-major (P = 0.00022, **C**), and low levels of LumSec-myo cells (P = 0.019, **D**) were significantly associated with shorter time to IBC recurrence.


A
Basal
LumSec
LumHR
Fibroblasts
Major Cell Type
ER Status
Case
log (Cell Count)
Up-Regulated Basement-Membrane DEGs
MMRN2
DST
ACTA2
TNC
COL7A1
COL9A3
COL8A2
LAMA1
P3H2
COL17A1
LAMC2
ITGB4
LAMB3
LAMA3
AGRN
CD151
COL4A5
COL4A6
VWA1
ENTPD2
ANXA2
PXDN
LAMB2
COLQ
COL4A3
COL4A4
TMEFF1
ANG
TIMP1
TIMP3
ADAMTS1
SPARC
COL4A1
COL4A2
HSPG2
SERPINF1
LAMC1
CCDC80
LAMA4
EFEMP2
EGFLAM
EGFL6
LOXL2
COL18A1
FN1
COL6A1
NID2
FREM1
HMCN1
SMOC2
LOXL1
FBLN1
LAMA2
FBN1
THBS2
COL5A1
LAMB1
NID1
COL15A1
TGFBI
LAMC3
THBS4
AFRN
NTNG2
FRAS1
NPNT
DLG1
LAMA5
NTN4
VWA2
MEGF9
LAD1
EFNA5
Major Cell Type
–log 10 (Adj. P)
−log10(Adj. p)
100
80
60
40
20
0
log(Cell Count)
10
8
6
4
2
Major Cell Type
Basal
LumSec
LumHR
Fibroblasts
Expression
4
2
0
−2
−4
Case
2
10
12
16
PCA−001
PCA−19−005
PCA−20−025
PCA−20−028
PCA−20−030
PCA−20−033
PCA−20−035
PCA−20−037
PCA−21−042
PCA−21−050
PCA−22−053


B
E
S
Basal
LumSec
LumHR
Fibroblast
COL4A1

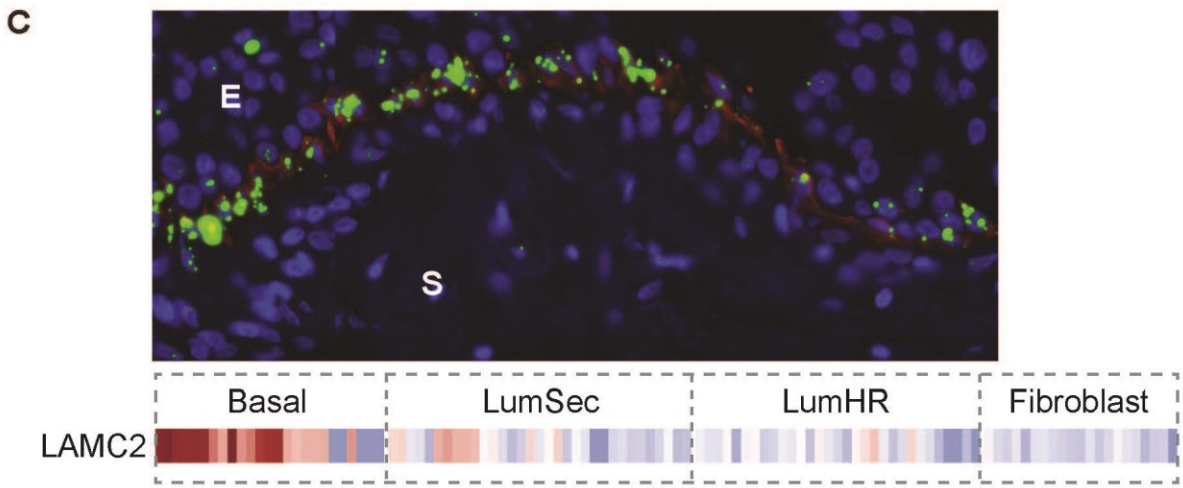
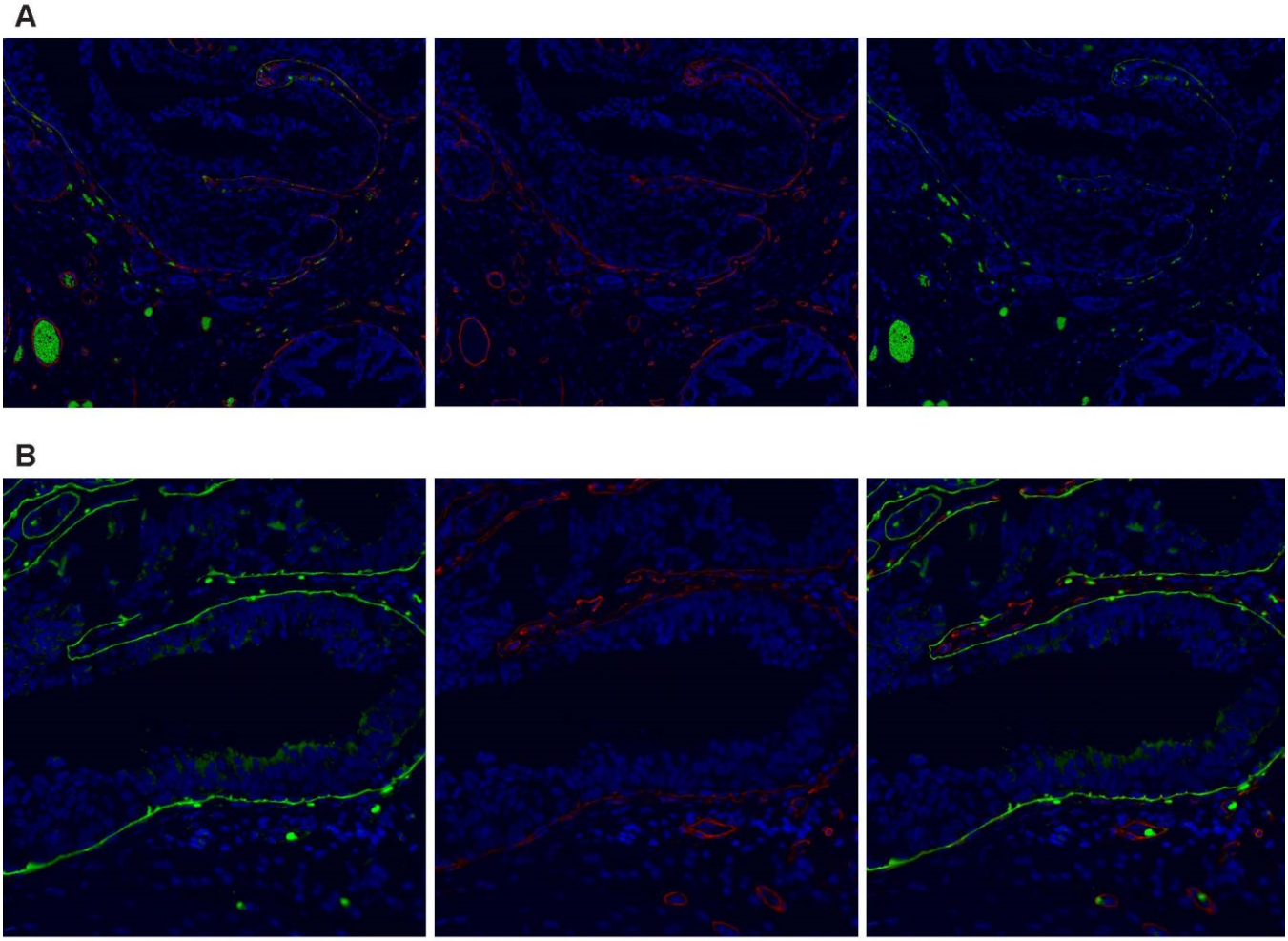
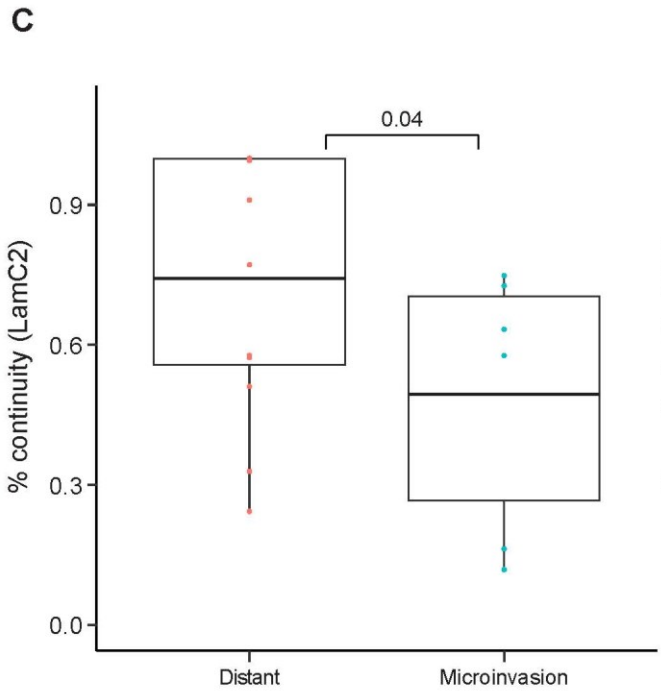

C
E
S
Basal
LumSec
LumHR
Fibroblast
LAMC2

Figure 5

Figure 5. Basement membrane gene expression from scRNA-seq.
A) Relative expression of basement membrane genes across three major epithelial cell types and fibroblasts (scRNA-seq data). Displayed genes were significantly up-regulated (adj. $p < 0.01$) given differential expression between the cell types. **B)** Detection of COL4A1 (green) in DCIS (E) and adjacent stroma (S) with ACTA2 protein expression demarcating the basal zone of the DCIS sample, with the corresponding gene expression of the Basal, LumSec, LumHR, and Fibroblasts below. **C)** Detection of LAMC2 (green) in DCIS (E) and adjacent stroma (S) with ACTA2 protein expression demarcating the basal zone of the DCIS sample, with the corresponding gene expression of the major epithelial cells and fibroblasts below (from panel A).





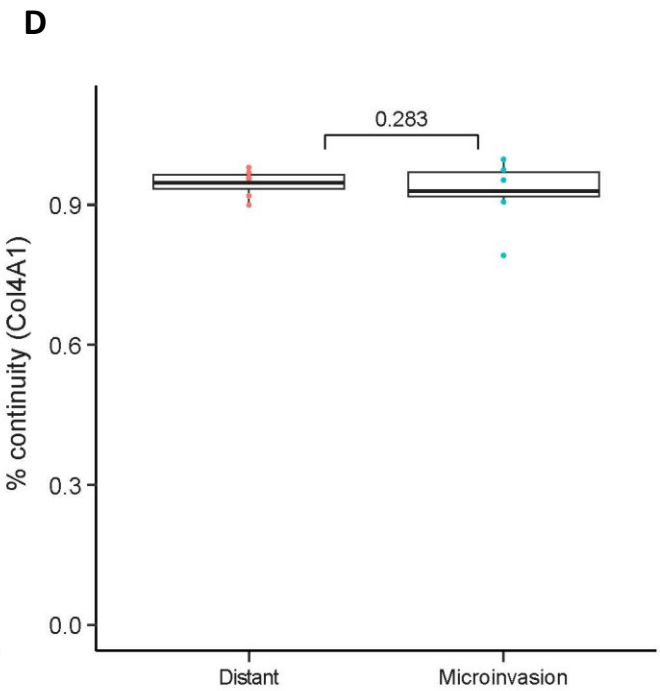


Figure 6. Basement membrane expression by in situ analysis.
A, **B)** Immunofluorescence of COL4A1 (**A**), LAMC2 (**B**) in DCIS located adjacent to microinvasion (Sample set 4 Microinvasive). **C**, **D)** Immunofluorescence of COL4A1 (**C**), LAMC2 (**D**) in DCIS located distant to microinvasion (Sample set 4 Microinvasive). **E, F)** Comparison of percentage of continuity for LAMC2 (**E**), COL4A1 (**F**) in DCIS ducts between DCIS adjacent to microinvasion or distant to microinvasion.

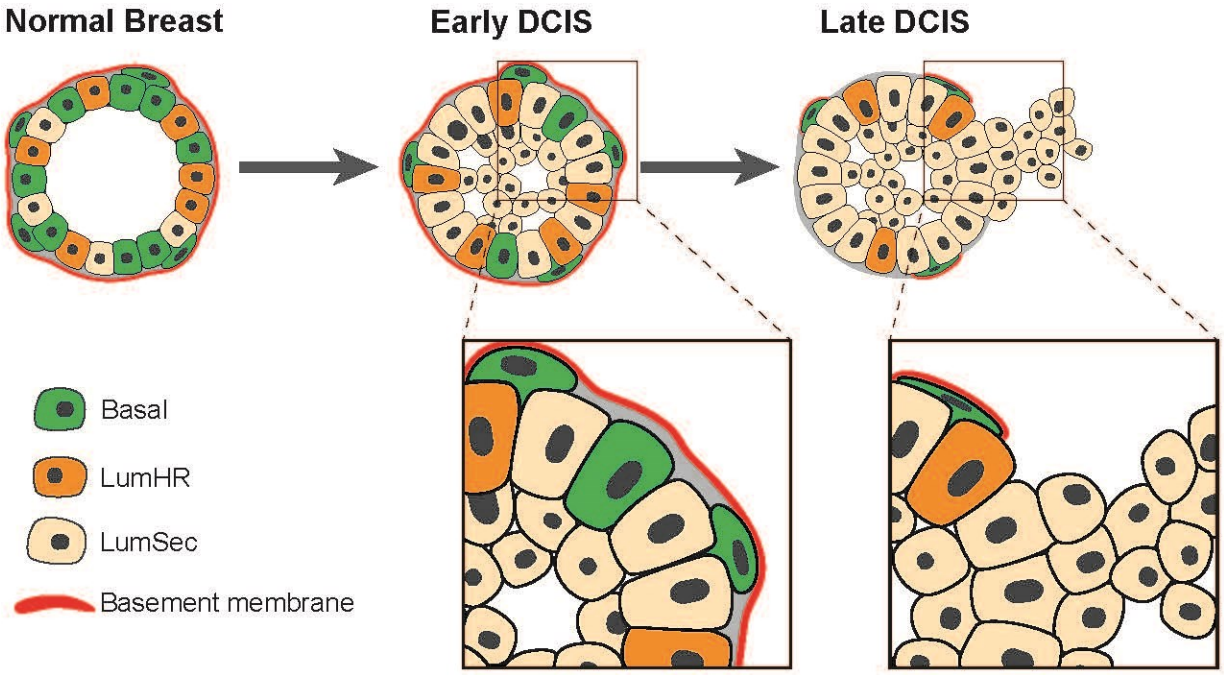


Figure 7. Conceptual model of DCIS invasion. Hypothesized changes in epithelial cell fractions during disease progression from normal breast to early and late DCIS stages.

Methods

Sample sets

Sample set 1 (TA578 dataset) consisted of DCIS (n=16), Invasive (n = 2) and synchronous normal tissues (n=12) from 16 patients undergoing mastectomy at Duke and Stanford, collected between 2019 – 2023. Due to low sample size, the invasive samples were not included the main analyses. Both the DCIS and normal specimens were bisected with one half used for single-cell processing, and the other half formalin fixed and embedded in paraffin (FFPE) (**Figure 1A**). FFPE samples were used for histologic examination to confirm the presence and extent of DCIS and normal breast epithelium. One of the samples was initially annotated as normal but upon further review was determined to be DCIS. The FFPE samples were also used for low-pass whole genome sequencing (WGS DNA-seq) (matched DCIS n = 8) and "bulk" RNA sequencing on microdissected areas of DCIS to capture epithelial tissue (LCM Bulk RNA-seq) (DCIS n = 12, Normal n = 3) (**Figure 1A**).

Sample set 2 consisted of 169 epithelial FFPE samples collected from 62 patients at Stanford Hospital 2008-2015. Among these patients, 42 had available ER status data and matched samples from Normal, DCIS, IBC tissues. Samples were dissected using LCM for epithelium followed by RNA-seq as described below.

Sample set 3 consisted of samples from the combined TBCRC and RAHBT cohorts from the HTAN DCIS Atlas[1]. We used RNA-seq data from macrodissected primary DCIS FFPE samples (TBCRC n = 216 samples, RABHT n = 97 samples). For the final analysis in this paper we included only samples from patients with either no recurrence (n = 163 patients) or IBC progression (n = 69 patients). For more information see [1].

Sample set 4 (Microinvasive dataset) consisted of 57 epithelial FFPE samples collected from 13 patients at Stanford Hospital 2006-2014. These patients had matched samples from Normal and DCIS. Samples were organized on a tissue microarray and the epithelial component dissected using LCM.

Sample set 5 consisted of 59 invasive prostate cancer, 46 prostate intraepithelial neoplasia (PIN), and 51 normal prostate FFPE samples collected at Stanford Hospital between 2019 and 2020. Samples were organized into a TMA and the epithelial component was dissected using LCM followed by RNA-seq as described below.

Single-cell RNA-sequencing (scRNA-Seq) assay

Fresh tissue samples were collected within 1 h 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 degrees C while awaiting pathological confirmation of DCIS in the matching DCIS FFPE

H&E 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 the Miltenyi Human Tumor Dissociation Kit, (130-095-929) according to protocol with gentleMACS Octo Dissociator with Heaters (130-096-427) program 37C_h_TDK_3. Cells were washed using 3 mL Gibco RPMI 1640 Medium (ThermoFisher Scientific cat# 11875119), strained using MACS smart strainer 70 um, and counted on a Countess II cell counter (ThermoFisher Scientific). Cells were resuspended in Hanks' Balanced Salt Solution (ThermoFisher Scientific Cat# 88284) to 12000 cells/uL for immediate downstream processing. GEMs 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 an Illumina NovaSeq.

Bulk-RNA sequencing assay

The bulk-RNA sequencing libraries were prepared for sample sets 1,2, 4, and 5 using LCM to capture epithelial tissue as described previously [1] and then prepared for sequencing using the SMART-3SEQ protocol [2]. Libraries were sequenced on an Illumina NextSeq 500 instrument with a High Output v2.5 reagent kit (Illumina # 20024906).

Low-pass whole-genome DNA-sequencing (WGS DNA-Seq) assay

Genomic DNA was isolated from FFPE samples using PicoPure DNA Extraction kit (Thermo Fisher Scientific # KIT0103). DNA libraries and sequencing were performed as previously published[1].

ImmunoFluorescence assay

ImmunoFluorescence (IF) staining was performed on paraffin-embedded Tissue Microarray (TMA) sections (4 µm) (Sample set 4 Microinvasive dataset). Briefly slides were deparaffinized and hydrated using a xylene-ethanol series. Antigen retrieval was carried out at 116ºC for 3 min. in a decloaking chamber in Antigen Unmasking Solution, Citrate-based pH6, at 10mM (Cat No. S236984-2 Agilent, Santa Clara, CA) using Col IV (Sigma Cat# AB769; 1:50) and SMA (ThermoFisher Cat# 53-9760-82; 1:3000) antibodies. For Lamc2 (SantaCrus Biotech Cat# sc-28330; 1:50), antibody proteinase K (Agilent Tech Cat# S302080-2) digestion was done. Slides were incubated with primary antibody at room temperature for 45 min. To detect Alexa Fluor conjugated secondary antibodies at 1:700 dilution was used. Counter staining was performed using Prolong Gold Antifade with DAPI (ThermoFisher Cat# P36935). Slides were scanned with a Leica Microscope slide scanner using Ariol Software (Leica Biosystems, CA).

Immunohistochemistry and RNA ISH co-detection assay

Co-detection of ActA2 (SMA) protein and COL4A1 and LAMC2 RNA were performed using all ACD guidelines for performing in situ hybridization with the RNAscope 2.5 HD Detection Kit-RED (Cat# 322360) and RNA-Protein Co-Detection Ancillary kit (Cat# 323180) immunochemistry on FFPE tissue sections (Sample set 4 Microinvasive dataset). The following RNA probes were used: RNAscope Probe - Hs-COL4A1 Cat#

461881 and RNAscope Probe - Hs-LAMC2 Cat# 501371. Immunohistochemistry 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.

Basement membrane continuity analysis

To evaluate the continuity of basement membrane in the IF images, the basement membrane location was annotated using Qupath [3]. The percentage of continuity of the basement membrane for each duct was determined by calculating the ratio of the total number of intact pixels to the total number of pixels in the basement membrane. Each duct was paired with a pathologist-annotated mask that indicates the precise location of the duct on the 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 regarding 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 if a particular location is deemed intact. This scanning process was conducted within a confidence range of 250-500 pixel units to account for potential variability in annotations.

WGS DNA-Seq data processing and analyses

DNA sequencing data were preprocessed using BWA-MEM algorithm v0.7.17 [4] for sequence alignment to the reference genome GRCh38/hg38 and GATK v4.1.7.0 [5-7] to mark duplicates and calibrate reads within the Nextflow-based pipeline Sarek v2.6.1 [8,9]. The recalibrated reads were further processed and filtered for mappability, GC content using the R/Bioconductor package QDNAseq v1.22.0 [10] with R statistical environment v3.6.0 [11]. For QDNAseq, 50-kb bin annotations were obtained from QDNAseq.hg38 (v1.0.0, https://github.com/asntech/QDNAseq.hg38). We kept only autosomal sequences after filtering due to low-depth mappability and GC correction. The R package ACE v1.4.0 [12] was used to estimate and identify the max cellularity for different ploidies (2, 3, 4). Copy number aberrations were called using CGHcall v2.48.0 [13].

Bulk RNA-seq data processing

The RNA sequencing libraries were pre-processed using 3SEQtools [2]. Read alignment was conducted using STAR v 2.7.3a [14]. Gene level reads were obtained using featurecounts Rsubread package v4.0.0 [15]. The reference sequence and genomic annotation files were obtained from GENCODE [16].

scRNA-Seq data processing and analyses

Preprocessing: To process the raw scRNA-Seq data from the 30 sequenced libraries into a cell by gene matrix with read counts, 10X Genomics Cell Ranger v6.0.1 [17] was used. Application of this pipeline yielded 196,848 filtered cells from 30 libraries. Briefly, the mkfastq module used bcl2fastq to demultiplex the reads, then the count module aligned and mapped 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 [18] using the Seurat package v4.3.0.1 [19] to perform quality control 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. Highly-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 [20] were prefixed by the string "MT-". At a minimum, cells were assumed to be 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.

A total of 8 libraries (3 DCIS and 5 normal) were excluded from downstream analyses due to an insufficient number of high-quality cells (< 2,000). One of the samples was initially annotated as normal but upon further review was determined to be DCIS. The remaining 22 libraries (7 normal, 13 DCIS and 2 invasive) containing 146,146 cells were derived from 15 patients: 1 patient with matched normal, DCIS and invasive libraries (3 libraries), 4 patients with paired normal and DCIS libraries (8 libraries), 7 patients with DCIS-only libraries (8 libraries), 2 patients with normal-only patient libraries (2 libraries), and 1 patient with an invasive-only library (1 library).

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 8 sequencing batches. Integration was performed using Seurat functions [19] that involve (i) merging sample count matrices by batch then (ii), within each batch counts are log-normalized and the top 2,000 most variable gene features are selected based on variance-stabilizing transformation. Next, (iii) variable gene features across batches, i.e. integration features, are selected from the union of top batch-wise variable features and these are used to reduce the dimensionality of each dataset and then identify mutual nearest neighbors, i.e. anchors. Integration returns 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 was scaled, reduced in dimension using Principal Component Analysis (PCA) on the top 30 components, and subsequently reduced onto 2-dimensional Uniform Manifold Approximation and Projection (UMAP) coordinates [21] using Seurat functions [22]. Unsupervised clustering of the cells were performed with Seurat functions [22] by constructing a shared nearest-neighbor (SNN) graph [23] based on the top 30 PCA

components and then optimizing the standard modularity function based on the Louvain algorithm [24] with the resolution parameter set to 0.5, as advised in Seurat tutorial.

Cell-type inference: To infer a cell type classification for each cell based on gene expression, the R package scSorter v0.0.2 [25] was used along with a curated set of signature genes. The curated signature gene set consisted of genes to profile the 10 major cell types in the normal breast tissue, including Basal, LumSec, LumHR, B cells, Fibroblasts, Lymphatic, Myeloid, Perivascular, T cells, Vascular (see Supplementary Table 3 from Kumar *et al.* [26]) and 11 epithelial cell states, including Basal, LumSec-basal, LumSec-HLA, LumSec-KIT, LumSec-lac, LumSec-major, LumSec-myo, LumSec-prol, LumHR-active, LumHR-major, LumHR-SCGB (see Supplementary Table 10 from Kumar *et al.* [26]). Specifically, the signature genes for the 3 major epithelial cell types in the former set were replaced by the signature genes for the 11 epithelial cell states 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 equal or less than 10% of cells. Counts were log-normalized and a scSorter [25] tuning parameter alpha of 0.2 was selected to allow for unknown cells and represent the most common composition of cell types across a range of alpha values.

Copy-number state inference: We distinguished the DCIS cells in DCIS libraries (DD) from the non-neoplastic epithelial cells in DCIS libraries (DN) based on their RNA expression-inferred CNV profile (**Figure 2B**, **Figure 3A**, **Figure 3D**) and co-located DNA-inferred CNV profiles (see **WGS DNA-Seq data processing** section above). Although the DNA derived CNV represents an average of cells in the DCIS sample, it enabled us to distinguish CNV related RNA expression changes from spurious non-CNV related coordinated local RNA expression changes. The inferCNV package v1.10.1 [27] was used to infer the copy number states of cells from DCIS samples using scRNA data. To this end, a panel of 51,151 normal epithelial cells consisting of epithelial cells derived from the 7 normal scRNA-seq libraries was constructed to estimate the normal state as a reference. A cutoff of 0.1 for the minimum average read counts per gene among reference cells, default settings for Hidden Markov Model (HMM) and de-noising filters were used. In addition, the analysis mode of “subclusters” was used to predict CNV at the levels of subpopulations for each scSorter-inferred cell type (see **Cell-type inference** section above).

Co-located DNA-inferred CNV profiles were available for eight out of the 14 patients in the final analysis data. The CNV profiles were initially assessed through a manual review process based on investigators' knowledge. During this process, specific DNA-inferred CNV regions with directionality (gain/loss) were identified to serve as decision rules for classifying each individual cell as either a neoplastic or non-neoplastic cell. For a given cell within a scRNA-seq sample, if the number of RNA-inferred CNV regions aligning with the decision rule regions exceeded a sample-specific threshold, the cell was designated as a neoplastic cell. In cases where scRNA-seq samples lacked a co-located DNA-

inferred CNV profile, the set of decision rule regions was derived from the CNV profile of the scRNA-seq sample itself. In total, 17,454 DD and 13,899 DN cells were identified in 13 DCIS libraries, 14,555 NN cells were identified in 7 normal libraries. The decision rules and CNV profiles for the 13 samples are provided in **Extended data Table 2 and Supplementary data file 1**.

Pseudo-time trajectory inference: The monocle package v2.26.0 [28-30] was used to infer the pseudo-time trajectory of epithelial cell expression. The read count matrix of the epithelial cells was filtered to remove rarely-expressed genes (i.e., excluded genes found in equal or less than 100 cells, or showing an expression value less than 0.1), and underwent dimension reduction to 2-dimensions using Discriminative Dimensionality Reduction with Trees (DDRTree) [31] and then cells were ordered in the reduced space using the signature genes used for cell-type inference. The root point of the trajectory was assigned to the Monocle-inferred state which contained the most Basal cells.

Pseudo-bulk analysis: To investigate differentially expressed genes between CNV-based groups and ER status (**Figure 2C**), raw cell-level counts were aggregated from each patient and CNV-based group, and then differential expression analysis and gene set enrichment analysis were performed for various contrasts to demonstrate the CNV-base approach can distinguish neoplastic and non-neoplastic epithelial cells (DD vs. DNNN, DD vs. NN, DD vs DN, DN vs NN) and explore the influence of ER status (DD ER+ vs. DNNN, DD ER- vs. DNNN, DD ER+ vs. DD ER-). To investigate differences in expression profiles between cell types (**Figure 5A**, **Supplemental Figure 3A**, **Supplemental Figure 5**), raw cell-level counts were aggregated from each patient and cell type, and then differential expression analysis was performed between a given cell type versus all others. Differential expression analysis was performed using the DESeq2 package v1.38.3 [32] in R. For Gene Set Enrichment Analysis, all genes were pre-ranked by the Wald statistics from the DESeq2 results prior to the enrichment analysis of Hallmark pathways using the fgsea package v1.24.0 [33]. The pathways were acquired from MSigDB [34,35] via the msigdbr package v7.5.1 [36]. All p-values were adjusted to control for multiple testing using the Benjamini-Hochberg procedure [37].

Cell type abundance imputation: The CIBERSORTx tool [38] was used to infer the cell type composition within bulk RNA-seq data sets (Sample sets 1-5), using public single-cell data sets with curated cell types as reference. For each bulk data set, its raw gene count matrix was used as the input, while an external scRNA-seq count matrix with original cell assignments was used to construct the cell type signature matrix. scRNA-seq data from normal breast samples [26] was used as the reference for breast bulk data sets (Sample set 1 TA578, Sample set 2 S3S, Sample set 3 DCIS Atlas, Sample set 4 Microinvasive), whereas the scRNAseq data from normal prostate samples [39] was used as the reference for the prostate bulk data set (Sample set 5 PIN). Both matrices were randomly down-sampled 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 scRNAseq-derived signature matrices, was applied to correct for cross-platform batch effects. For each sample set, the cell fractions were imputed for all samples in Sample set 1-5, but the stroma only samples in Sample sets 1-2, 5 were excluded from downstream analysis, since our focus was on the epithelial tissue. The raw imputed cell fractions from 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 scRNASeq data, the Wilcoxon rank-sum test was used. In the case where bulk data had a patient-matched design (Sample set 1 TA578), 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 [37] (**Figure 3B**, **3E**, **4A**). Fisher's exact test was used to compare the overall fractions of cells of each contrast group, within each cell type (**Figure 3C**, **3F**). Pearson's correlation test was used to evaluate the correlation between the cell type abundance estimated from the scRNASeq data and the corresponding bulk data (Sample set 1 TA578) on a per-sample basis (**Extended data Figure 4A**).

Outcome analysis

Patients in sample set 3 were divided into ‘high’ and ‘low’ cell fraction groups based on the median of each respective cell type. Associations with time to IBC recurrence were quantified using Cox Proportional Hazard model [40]. Kaplan-Meier plots as implemented in the R packages survival v3.3-1 [40] and survminer v0.4.9 [41] were used to visualize outcome differences.

Data availability

The scRNA sequencing data, WGS data, bulk RNA sequencing datasets and imaging data have been deposited via the Human Tumor Atlas Network (HTAN) dbGaP Study

Accession: phs002371.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002371.v1.p1).

Code availability

The scripts to reproduce the analyses presented in this paper are available publicly through a gitlab repository (https://).

1. Strand, S.H. *et al.* Molecular classification and biomarkers of clinical outcome in breast ductal carcinoma in situ: Analysis of TBCRC 038 and RAHBT cohorts. *Cancer Cell* **40**, 1521-1536 e7 (2022).
2. Foley, J.W. *et al.* Gene expression profiling of single cells from archival tissue with laser-capture microdissection and Smart-3SEQ. *Genome Res* **29**, 1816-1825 (2019).
3. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci Rep* **7**, 16878 (2017).
4. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
5. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-8 (2011).
6. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 1-33 (2013).
7. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
8. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35**, 316-319 (2017).
9. Ewels, P.A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* **38**, 276-278 (2020).
10. Scheinin, I. *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* **24**, 2022-32 (2014).
11. R Core Team. R: A Language and Environment for Statistical Computing. *https://www.R-project.org/* (2022).
12. Poell, J.B. *et al.* ACE: absolute copy number estimation from low-coverage whole-genome sequencing data. *Bioinformatics* **35**, 2847-2849 (2019).
13. van de Wiel, M.A. *et al.* CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics* **23**, 892-4 (2007).
14. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
15. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-30 (2014).
16. Frankish, A. *et al.* Gencode 2021. *Nucleic Acids Res* **49**, D916-D923 (2021).
17. Zheng, G.X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).
18. Team, R.C. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. *http://www. R-project. org/* (2013).
19. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587. e29 (2021).
20. Tweedie, S. *et al.* Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res* **49**, D939-D946 (2021).
21. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
22. Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888-1902. e21 (2019).
23. Houle, M.E., Kriegel, H.P., Kroger, P., Schubert, E. & Zimek, A. Can Shared-Neighbor Distances Defeat the Curse of Dimensionality? *Scientific and Statistical Database Management* **6187**, 482-+ (2010).
24. Waltman, L. & Van Eck, N.J. A smart local moving algorithm for large-scale modularity-based community detection. *The European physical journal B* **86**, 1-14 (2013).
25. Guo, H. & Li, J. scSorter: assigning cells to known cell types according to marker genes. *Genome Biol* **22**, 69 (2021).

26. Kumar, T. *et al.* A spatially resolved single-cell genomic atlas of the adult human breast. *Nature* **620**, 181-191 (2023).
27. Tickle, T., Tirosh, I., Georgescu, C., Brown, M. & Haas, B. inferCNV of the Trinity CTAT Project. *Klarman Cell Observatory, Broad Institute of MIT and Harvard* (2019).
28. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386 (2014).
29. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat Methods* **14**, 309-315 (2017).
30. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979-982 (2017).
31. Mao, Q., Wang, L., Tsang, I.W. & Sun, Y. Principal graph and structure learning based on reversed graph embedding. *IEEE transactions on pattern analysis and machine intelligence* **39**, 2227-2241 (2016).
32. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
33. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv*, 060012 (2021).
34. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
35. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425 (2015).
36. Dolgalev, I. msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format. R package version 7.2. 1. *[Internet]* (2020).
37. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **57**, 289-300 (1995).
38. Newman, A.M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology* **37**, 773-782 (2019).
39. Henry, G.H. *et al.* A Cellular Anatomy of the Normal Adult Human Prostate and Prostatic Urethra. *Cell Rep* **25**, 3530-3542 e5 (2018).
40. Therneau, T.M. & Grambsch, P.M. *Modeling survival data : extending the Cox model*, xiii, 350 p. (Springer, New York, 2000).
41. Kassambara, A., Kosinski, M., Biecek, P. & Fabian, S. survminer: Drawing Survival Curves using 'ggplot2'. *R package version 0.3* **1**(2017).
42. Strand, S.H. *et al.* Molecular classification and biomarkers of clinical outcome in breast ductal carcinoma in situ: Analysis of TBCRC 038 and RAHBT cohorts. *Cancer Cell* **41**, 1381 (2023).

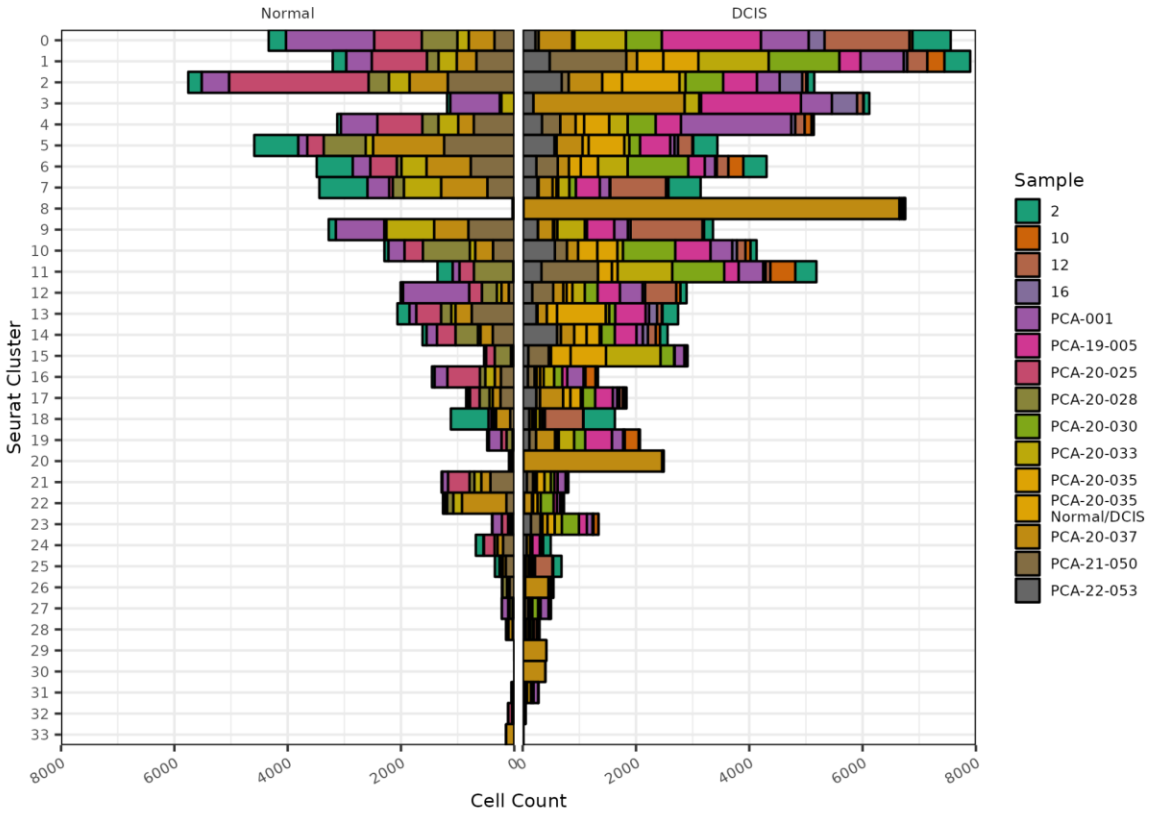
Extended data table 1. Baseline demographic and clinical characteristics for DCIS scRNA-seq patient study cohort.

Batch	scRNA Patient ID	scRNA Sample ID	scRNA Sample Type	scRNA Sample Type (Merged)	ER Status	Post-QC Cell Count	Epithelial Cell Count	Matched WGS Sample ID	Matched Smart-3Seq Sample IDs
Duke_Set1	PCA-001	PCA-001-L	DCIS	DCIS	+	7886	2153	TA578_L13	14236_592_DCIS_Ep
Duke_Set1	PCA-001	PCA-001-M	Normal	Normal	NA	8681	3812		
Duke_Set2	PCA-19-005	PCA-19-005T	DCIS	DCIS	+	10310	4886		14201_574_NL_Ep
Duke_Set2	PCA-20-025	PCA-20-025N	Normal	Normal	NA	9359	1086		
Duke_Set2	PCA-20-028	PCA-20-028N	Normal	Normal	NA	6066	1432		
Duke_Set3	PCA-20-030	PCA-20-030-DCIS	DCIS	DCIS	-	8626	1773	TA578_L9	14210_579_DCIS_Ep
Duke_Set3	PCA-20-033	PCA-20-033-N	Normal	Normal	NA	4455	1949		
Duke_Set3	PCA-20-033	PCA-20-033-T	DCIS	DCIS	-	7396	2747	TA578_L10	14220_584_DCIS_Ep
Duke_Set4	PCA-20-035	PCA-20-035-DCIS	DCIS	DCIS	-	6197	258	TA578_L11	14230_589_DCIS_Ep
Duke_Set4	PCA-20-035	PCA-20-035-Normal_Flagged	Normal/DCIS	DCIS	-	3235	384		14230_589_DCIS_Ep
Duke_Set4	PCA-20-037	PCA-20-037-DCIS	DCIS	DCIS	+	17841	8142	TA578_L12	14240_594_DCIS_Ep
Duke_Set4	PCA-20-037	PCA-20-037-Normal	Normal	Normal	NA	8082	2119		
Duke_Set5	PCA-21-050	PCA-21-050-N	Normal	Normal	NA	9482	1830		
Duke_Set5	PCA-21-050	PCA-21-050-T	DCIS	DCIS	-	5030	1025		
Duke_Set6	PCA-22-053	PCA-22-053	DCIS	DCIS	-	6303	1480		
Stanford_Set1	2	SS02-01_DCIS	DCIS	DCIS	+	4983	2571	TA578_L1	14205_576_NL_Ep
Stanford_Set1	2	SS02-03_N	Normal	Normal	NA	5026	2327		14205_576_NL_Ep
Stanford_Set2	10	SS04-02_4D2	DCIS	DCIS	-	2194	783	TA578_L3	14222_585_DCIS_Ep, 14226_587_DCIS_Ep
Stanford_Set2	12	SS05-03_5D	DCIS	DCIS	+	7161	4381		14224_586_DCIS_Ep, 14234_591_DCIS_Ep
Stanford_Set2	16	SS05-07_7D	DCIS	DCIS	+	2009	770	TA578_L8	14209_578_DCIS_Ep, 14214_581_DCIS_Ep

Extended data table 2. The decision rules and CNV profiles for the 13 scRNAseq samples.

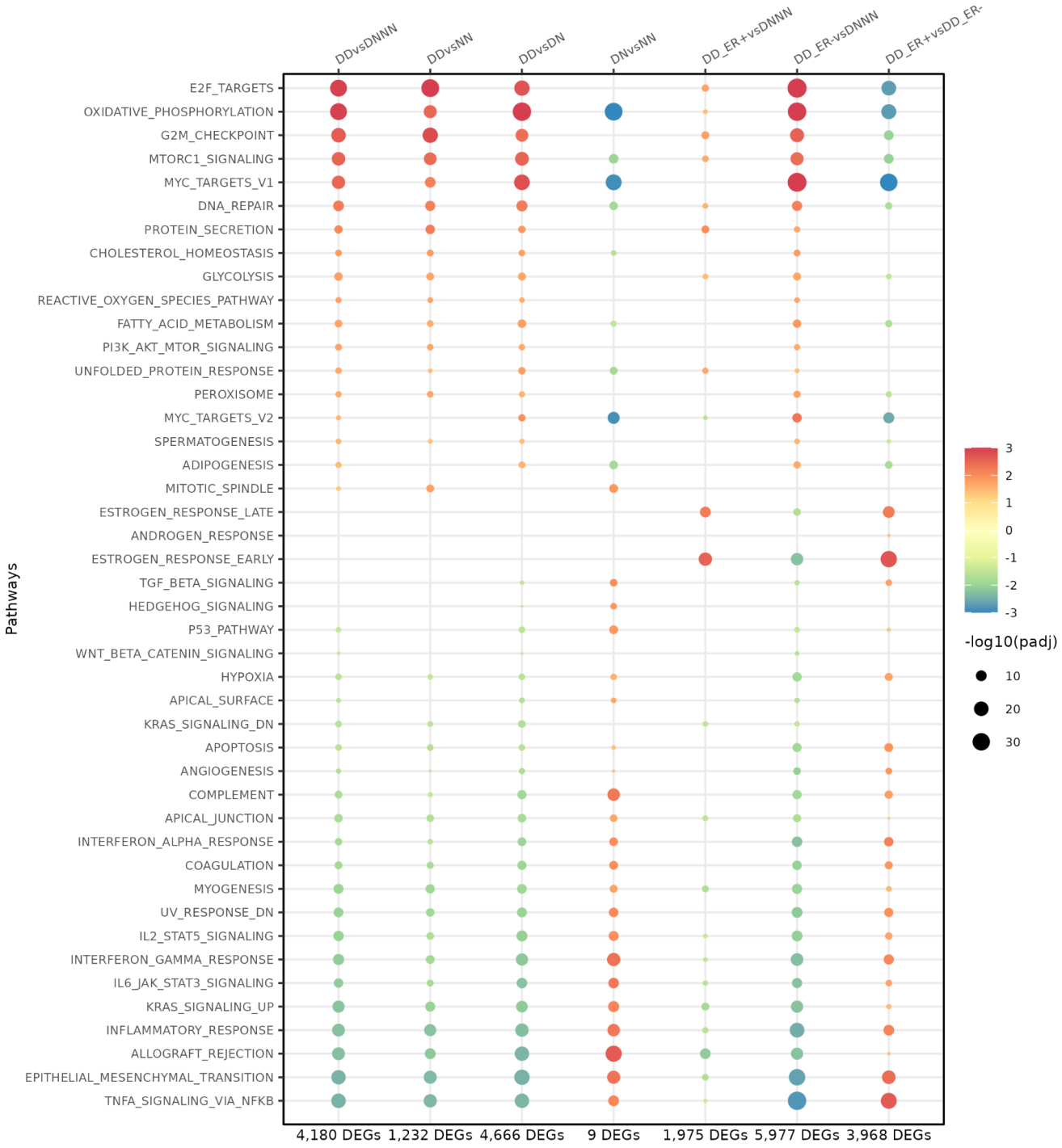
scRNA Patient ID	Matched WGS Sample ID	Chromosome	Start	End	Direction	Decision Rule
PCA-001	TA578_L13	chr1	143600001	171500000	1	2 out of 5 = DCIS
PCA-001	TA578_L13	chr1	171500001	225500000	1	2 out of 5 = DCIS
PCA-001	TA578_L13	chr3	50001	65100000	-1	2 out of 5 = DCIS
PCA-001	TA578_L13	chr13	18850001	88850000	-1	2 out of 5 = DCIS
PCA-001	TA578_L13	chr16	1	17000000	1	2 out of 5 = DCIS
10	TA578_L3	chr1	145700001	169850000	1	2 out of 5 = DCIS
10	TA578_L3	chr1	169850001	248900000	1	2 out of 5 = DCIS
10	TA578_L3	chr10	100001	38150000	1	2 out of 5 = DCIS
10	TA578_L3	chr11	36300001	48250000	1	2 out of 5 = DCIS
10	TA578_L3	chr19	37050001	40950000	1	2 out of 5 = DCIS
16	TA578_L8	chr8	32050001	41100000	1	1 out of 3 = DCIS
16	TA578_L8	chr20	35100001	44100000	1	1 out of 3 = DCIS
16	TA578_L8	chr20	52150001	59100000	1	1 out of 3 = DCIS
PCA-20-035	TA578_L11	chr1	187900001	228600000	1	2 out of 4 = DCIS
PCA-20-035	TA578_L11	chr4	52150001	189900000	1	2 out of 4 = DCIS
PCA-20-035	TA578_L11	chr8	86150001	100150000	1	2 out of 4 = DCIS
PCA-20-035	TA578_L11	chr17	74400001	79500000	1	2 out of 4 = DCIS
2	TA578_L1	chr1	61050001	247900000	1	2 out of 3 = DCIS
2	TA578_L1	chr16	1	31950000	1	2 out of 3 = DCIS
2	TA578_L1	chr17	150001	47950000	-1	2 out of 3 = DCIS
PCA-20-033	TA578_L10	chr3	42550001	93900000	-1	2 out of 7 = DCIS
PCA-20-033	TA578_L10	chr8	33900001	42050000	1	2 out of 7 = DCIS
PCA-20-033	TA578_L10	chr8	56300001	145100000	1	2 out of 7 = DCIS
PCA-20-033	TA578_L10	chr17	31050001	31450000	1	2 out of 7 = DCIS
PCA-20-033	TA578_L10	chr17	32300001	33050000	1	2 out of 7 = DCIS
PCA-20-033	TA578_L10	chr17	34400001	39850000	1	2 out of 7 = DCIS
PCA-20-033	TA578_L10	chr17	51800001	83200000	1	2 out of 7 = DCIS
PCA-20-037	TA578_L12	chr11	48200001	65500000	-1	2 out of 3 = DCIS
PCA-20-037	TA578_L12	chr17	40000001	48250000	-1	2 out of 3 = DCIS
PCA-20-037	TA578_L12	chr17	48250001	68050000	1	2 out of 3 = DCIS
PCA-20-030	TA578_L9	chr4	78650001	139050000	-1	1 out of 4 = DCIS
PCA-20-030	TA578_L9	chr5	50300001	171650000	-1	1 out of 4 = DCIS
PCA-20-030	TA578_L9	chr17	28450001	35100000	1	1 out of 4 = DCIS
PCA-20-030	TA578_L9	chr17	38350001	41850000	1	1 out of 4 = DCIS
12		chr11	108505435	134265855	-1	1 out of 3 = DCIS
12		chr16	69762306	90044975	-1	1 out of 3 = DCIS
12		chr22	30969245	44737681	-1	1 out of 3 = DCIS
PCA-19-005		chr11	60914158	66627347	-1	1 out of 4 = DCIS
PCA-19-005		chr11	102317450	134265855	-1	1 out of 4 = DCIS
PCA-19-005		chr17	43952738	83095122	1	1 out of 4 = DCIS
PCA-19-005		chr20	45361937	64287821	1	1 out of 4 = DCIS
PCA-21-050		chr11	167784	36659290	-1	1 out of 3 = DCIS
PCA-21-050		chr14	20256558	105856218	-1	1 out of 3 = DCIS
PCA-21-050		chr17	30378927	42615238	1	1 out of 3 = DCIS
PCA-22-053		chr9	15163622	33473930	1	1 out of 2 = DCIS
PCA-22-053		chr15	22838641	101724473	-1	1 out of 2 = DCIS

Extended data figure 1



Extended data Figure 1. Seurat clusters from normal and DCIS samples.
Cell counts for both normal breast (left side) and DCIS (right side) of Seurat clusters, color coded by each patient sample's contribution. Seurat cluster numbers correspond to those shown in Figure 2A.

Extended data figure 2

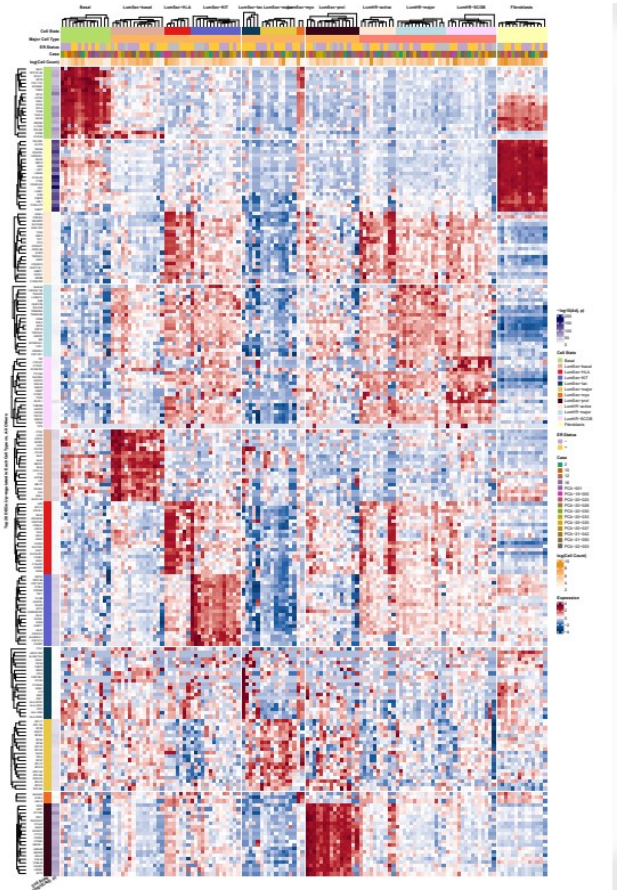


Extended data Figure 2. Gene Set Enrichment Analysis of neoplastic and non-neoplastic epithelial cell differences.

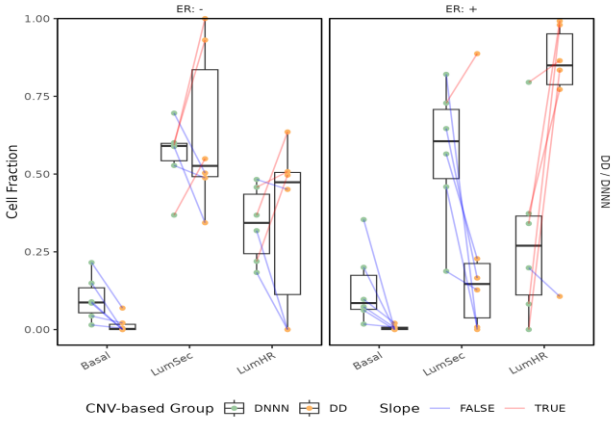
Gene Set Enrichment Analysis (GSEA) of all 50 Hallmark pathways between pseudo-bulk samples from neoplastic and non-neoplastic epithelial cells from breast tissue samples. The numbers of significant genes are indicated beneath the heatmap of the different comparisons. The color of the dots indicates the normalized enrichment score and the size indicates significance with the negative log10 adjusted *P*-value.

Extended data figure 3

A



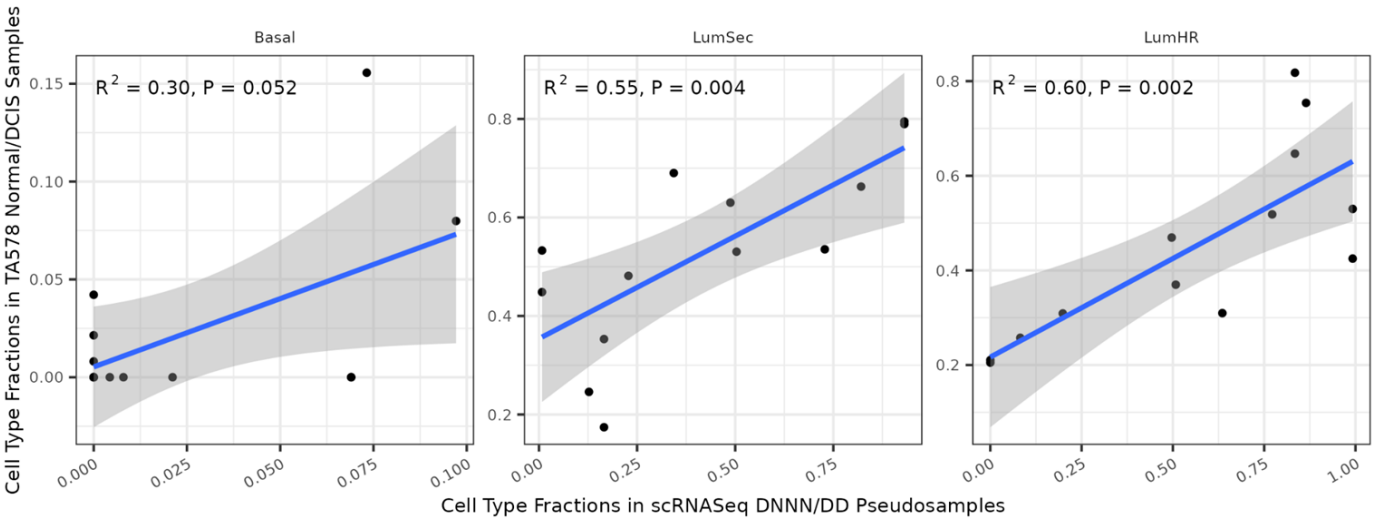
B



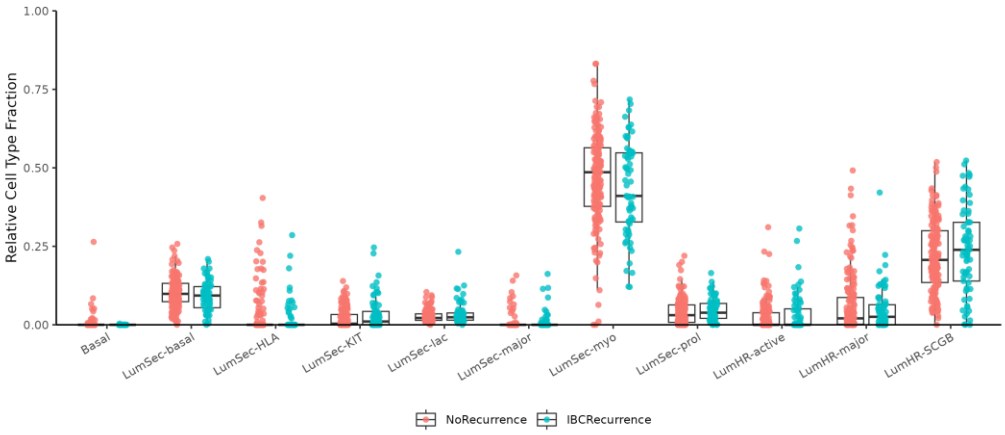
Extended data Figure 3. Cell substate expression and relative cell fractions.
A) Relative expression of the top 20 most up-regulated genes for each of the following cell types: 11 epithelial substates and fibroblasts, vascular, and perivascular cells. The displayed genes were determined based on differential expression of pseudo-bulked samples shown in columns; the top 20 up-regulated differentially expressed genes for each cell type are shown. Color bars indicate cell substate, cell state, CNV-based group (DD, DN, NN), ER status (DD samples only), patient ID, and cell count (natural log scale). **B)** Analysis of relative cell fractions based on 3 major epithelial cell types found in normal breast from scRNA-seq samples of 12 patient-matched normal breast, DCIS libraries (6 DD ER-, 6 DD ER+, and 12 normal; Sample set 1 TA578 dataset). Observed increases (decreases) are colored in red (blue). Pairwise differences were tested using one-sample Wilcoxon signed-rank tests. *P*-values were adjusted using Benjamini-Hochberg method to account for 6 comparisons.

Extended data figure 4

A

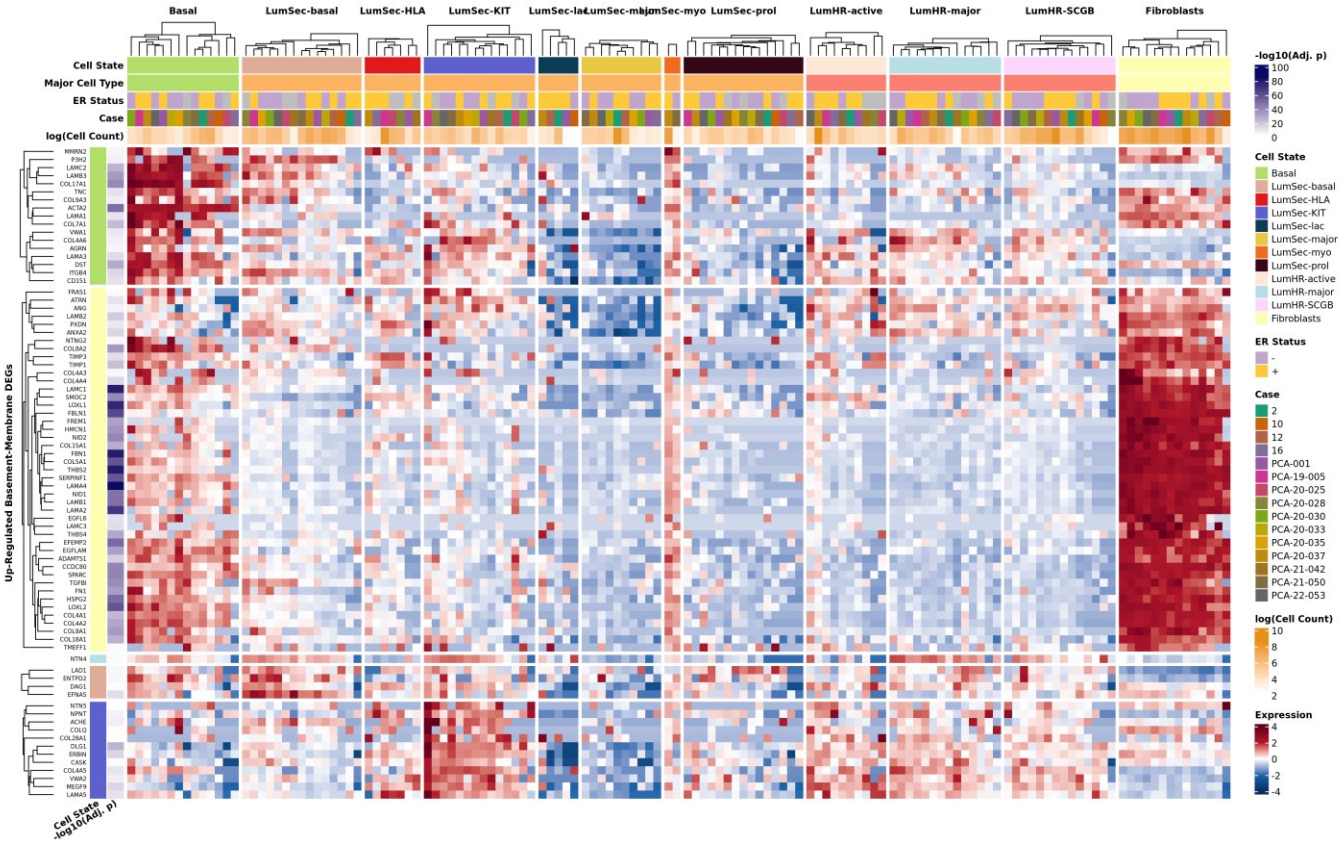


B



Extended data Figure 4. Bulk sample cell fraction analysis.
A) Correlation between relative cell fractions derived from the 10 scRNA-seq samples (4 DD ER-, 4 DD ER+, and 2 DNNN) and imputed relative cell fractions from 13 patient-matched FFPE bulk samples (5 ER- DCIS, 6 ER+ DCIS, and 2 normal), across 3 major epithelial cell types (Sample set 1 TA578 dataset). **B)** Analysis of imputed cell fractions based on 11 epithelial substates found in normal breast from Sample set 3 DCIS Atlas dataset with 163 DCIS_Only, and 69 DCIS_primary samples.

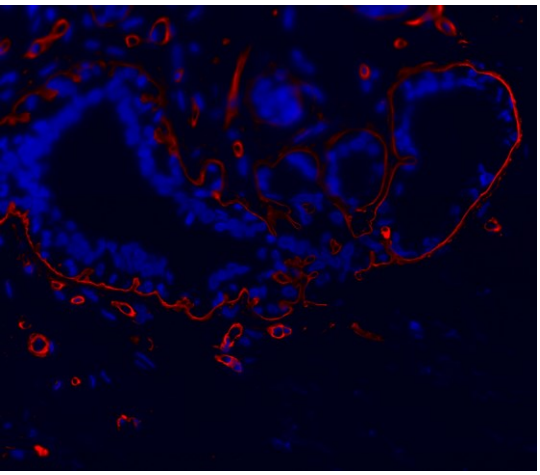
Extended data figure 5



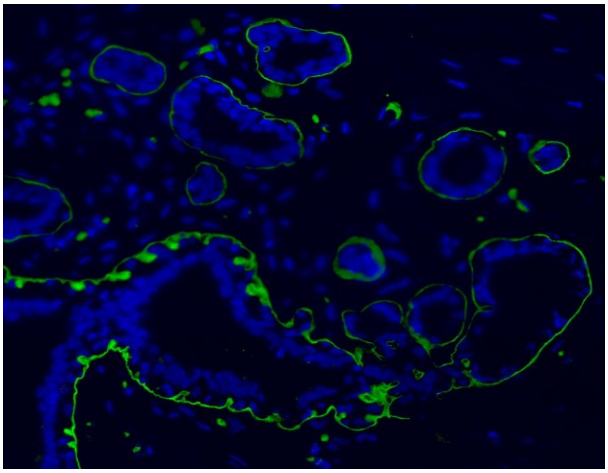
Extended data Figure 5. Relative expression of basement membrane genes across 11 epithelial cell substates. Relative expression of basement membrane genes across 11 epithelial cell substates and fibroblasts within each pseudo-bulk sample. The genes that are displayed were selected based on those that were significantly up-regulated (adj. $p < 0.01$) given differential expression between the cell types (Sample set 1 TA578 dataset).

Extended data figure 6

A

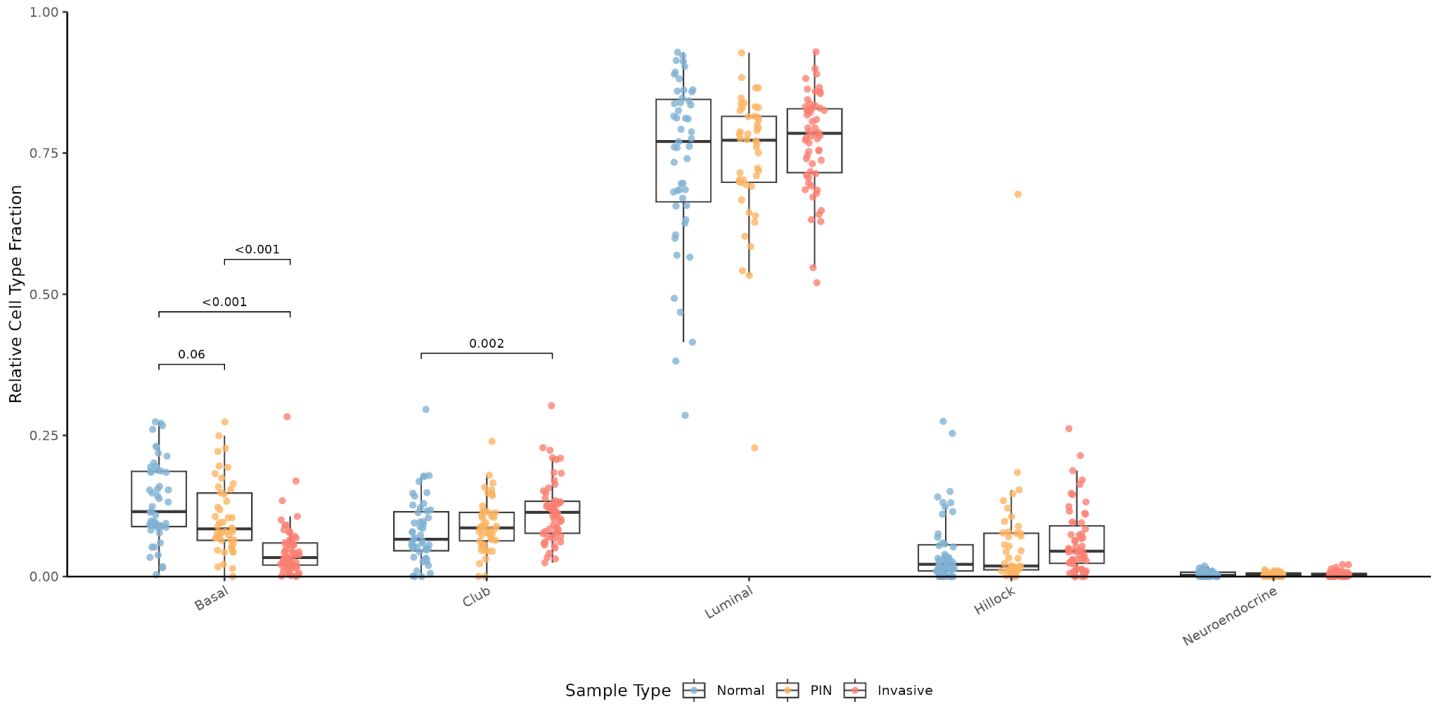


B



Extended data Figure 6. Immunofluorescence in normal breast.
A, B) Immunofluorescence of COL4A1 (**A**), LAMC2 (**B**) in normal breast epithelium (Sample set 4 Microinvasive).

Extended data figure 7



Supplemental Figure 7. Relative cell fractions in prostate tissues.
Relative cell fractions of 5 prostate epithelial cell types (Basal, Club, Luminal, Hillock, and Neuroendocrine) defined in normal prostate [39] from bulk-RNA sequencing datasets of the FFPE sample set with 51 normal prostate, 46 prostatic intraepithelium (PIN), and 59 invasive prostatic carcinoma (Sample set 5 PIN dataset). Differences in cell fraction by group were tested using Wilcoxon rank-sum and *P*-values were adjusted using Benjamini-Hochberg method to account for 15 multiple comparisons. The result comparing Basal between PIN and normal, and the results with an adjusted *P*-value of less than 0.05 are shown.