Open access • Posted Content • DOI:10.1101/2021.06.16.448585

DCIS genomic signatures define biology and correlate with clinical outcome: a Human Tumor Atlas Network (HTAN) analysis of TBCRC 038 and RAHBT cohorts

Siri H Strand, Siri H Strand, Belén Rivero-Gutiérrez, Kathleen E. Houlahan ...+50 more authors

Institutions: Aarhus University Hospital, Stanford University, Duke University, Arizona State University ...+13 more institutions

Published on: 24 Jul 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Breast cancer

DCIS genomic signatures define biology and correlate with clinical outcome: a Human Tumor Atlas Network (HTAN) analysis of TBCRC 038 and RAHBT cohorts

Siri H Strand[1,2], Belén Rivero-Gutiérrez[1#], Kathleen E Houlahan[3#], Jose A Seoane[3], Lorraine King[4], Tyler Risom[1], Lunden A Simpson[4], Sujay Vennam[1], Aziz Khan[3], Luis Cisneros[5], Timothy Hardman[4], Bryan Harmon[6,7], Fergus Couch[7,8], Kristalyn Gallagher[7,9], Mark Kilgore[7,10], Shi Wei[7,11], Angela DeMichele[7,12], Tari King[7,13,14], Priscilla F McAuliffe[7,15], Julie Nangia[7,16], Joanna Lee[7,17], Jennifer Tseng[7,18], Anna Maria Storniolo[7,19], Alastair Thompson[7,20], Gaorav Gupta[7,21], Robyn Burns[7,22], Deborah J Veis[23,24], Katherine DeSchryver[24], Chunfang Zhu[1], Magdalena Matusiak[1], Jason Wang[1], Shirley X Zhu[1], Jen Tappenden[25], Daisy Yi Ding[26], Dadong Zhang[27], Jingqin Luo[25], Shu Jiang[25], Sushama Varma[1], Lauren Anderson[4], Cody Straub[4], Sucheta Srivastava[1], Christina Curtis[3,28], Rob Tibshirani[26,29], Robert Michael Angelo[1], Allison Hall[30], Kouros Owzar[31], Kornelia Polyak[32], Carlo Maley[5], Jeffrey R Marks[4], Graham A Colditz[25], E Shelley Hwang[4]*, Robert B West[1]*

1. Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA
2. Department of Molecular Medicine, Aarhus University Hospital, 8200 Aarhus N, Denmark
3. Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA 94305, USA
4. Department of Surgery, Duke University School of Medicine, Durham, NC 27708, USA
5. School of Life Sciences, Arizona State University, Tempe, AZ 85281, USA
6. Department of Pathology, Montefiore Medical Center, Bronx, NY 10467, USA
7. TBCRC Loco-Regional Working Group
8. Department of Pathology, Mayo Clinic, Rochester, MN 55902, USA
9. Department of Surgery, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
10. Department of Pathology, University of Washington, Seattle, WA 98195, USA
11. Department of Pathology, University of Alabama at Birmingham, Birmingham, AL 35294, USA
12. Department of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
13. Breast Oncology Program, Dana-Farber Cancer Institute, Boston, MA 02215, USA
14. Department of Surgery, Brigham and Women's Hospital, Boston, MA 02115, USA
15. Department of Surgery, University of Pittsburgh, Pittsburgh, PA 15213, USA
16. Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston TX 77030, USA
17. Department of Surgery, MD Anderson Cancer Center, Houston, TX 77030, USA
18. Department of Surgery, University of Chicago, Chicago, IL 60637, USA
19. Department of Medicine, Indiana University, Indianapolis, IN 46202, USA
20. Department of Surgery, Baylor College of Medicine, Houston, TX 77030, USA
21. Department of Radiation and Oncology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
22. TBCRC, The EMMES Corporation, Rockville, MD 20850, USA
23. Department of Medicine, Washington University School of Medicine, St. Louis, MO 63108, USA
24. Departments of Pathology & Immunology, Washington University School of Medicine, St. Louis, MO 63108, USA
25. Department of Surgery, Washington University School of Medicine, St. Louis, MO 63110, USA
26. Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA
27. Duke Cancer Institute, Duke University School of Medicine, Durham, NC 27708, USA
28. Department of Medicine and Genetics, Stanford University, Stanford, CA 94305, USA
29. Department of Statistics, Stanford University, Stanford, CA 94305, USA
30. Department of Pathology, Duke University School of Medicine, Durham, NC 27708, USA
31. Department of Biostatistics & Bioinformatics, Duke University School of Medicine, Durham, NC 27708, USA
32. Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

[#] These authors contributed equally

*Correspondence: rbwest@stanford.edu, shelley.hwang@duke.edu

HIGHLIGHTS

- New transcriptomic classification solution reveals 3 major subgroups in DCIS.
- Four stroma-specific signatures identified.
- Outcome analysis identifies pathways involved in DCIS progression.
- CNAs characterize high risk of distant relapse IBC subtypes observed in DCIS.

SUMMARY

Ductal carcinoma *in situ* (DCIS) is the most common precursor of invasive breast cancer (IBC), with variable propensity for progression. We have performed the first multiscale, integrated profiling of DCIS with clinical outcomes by analyzing 677 DCIS samples from 481 patients with 7.1 years median follow-up from the Translational Breast Cancer Research Consortium (TBCRC) 038 study and the Resource of Archival Breast Tissue (RAHBT) cohorts. We made observations on DNA, RNA, and protein expression, and generated a *de novo* clustering scheme for DCIS that represents a fundamental transcriptomic organization at this early stage of breast neoplasia. Distinct stromal expression patterns and immune cell compositions were identified. We found RNA expression patterns that correlate with later events. Our multiscale approach employed *in situ* methods to generate a spatially resolved atlas of breast precancers, where complementary modalities can be directly compared and correlated with conventional pathology findings, disease states, and clinical outcome.

KEYWORDS

Ductal carcinoma in situ, RNA gene expression profiling, whole genome sequencing, multiplex immunohistochemistry, invasive breast cancer, precancer, outcome, human tumor atlas network, breast, tumor microenvironment.

INTRODUCTION

As nonobligate precursors of invasive disease, precancers provide a unique vantage point from which to study the molecular pathways and evolutionary dynamics that lead to the development of life-threatening cancers. Breast ductal carcinoma in situ (DCIS) is one of the most common precancers across all tissues, with almost 50,000 women diagnosed each year in the U.S. alone (American Cancer Society, 2019). Current treatment of DCIS involves surgical excision with either breast conserving surgery or mastectomy, with the goal of preventing invasive cancer. However, DCIS consists of a molecularly heterogeneous group of lesions, with highly variable risk of invasive progression. An improved understanding of which DCIS is likely to progress could thus spare a subgroup of women unnecessary treatment.

Identification of factors associated with disease progression has been the subject of substantial study. Epidemiologic models of cancer progression indicate that clinical features such as age at diagnosis, tumor grade, and hormone receptor expression may have some prognostic value; however, they have limited ability to identify the biologic conditions that govern whether DCIS will progress to invasive cancer. Previous molecular analyses of DCIS have studied either 1) cohorts of DCIS with known outcomes (*e.g.* disease-free versus recurrent), or 2) cross-sectional cohorts of DCIS that either do or do not exhibit adjacent areas of invasive cancer. Both of these approaches have tested key potentially divergent assumptions: Recurrence of the DCIS as IBC may arise from neoplastic cells that were left behind when the DCIS was removed, be related to an initial field effect, or may develop from independent events. Longitudinal cohorts have provided a perspective of cancer progression over time. Analysis of DCIS found adjacent to invasive cancer assumes that these preinvasive areas are a good model for pure DCIS tumors and are the ancestors of the invasive cancer cells, with synchronous lesions inferring progression. In either case, these studies have not produced clear evidence for a common set of events that are associated with invasion (Allinen et al., 2004, Gil Del Alcazar et al., 2017, Heselmeyer-Haddad et al., 2012, Lesurf et al., 2016, Newburger et al., 2013, Gorringe et al., 2015, Casasent et al., 2018, Abba et al., 2015, Vincent-Salomon et al., 2008a).

Lessons can be learned from precancerous evolution in other tissues. In Barrett's esophagus, the genomic copy number landscape and chromosomal instability predicted esophageal

cancer years before diagnosis (Killcoyne et al., 2020). In prostate cancer, a stromal signature reflecting immune and osteoblast activity stratified indolent from clinically significant disease (Tyekucheva et al., 2017), highlighting the relevance of the tumor microenvironmental context. These findings suggest diverse trajectories of premalignant to malignant tumor progression. This diversity is mirrored in DCIS where few genomic aberrations have been identified that can differentiate DCIS from IBC (Johnson et al., 2012, Heselmeyer-Haddad et al., 2012, Newburger et al., 2013, Gorringe et al., 2015, Yao et al., 2006, Pareja et al., 2020) and microenvironmental processes, including collagen organization, myoepithelial changes, and immune suppression, may contribute to IBC development (Lesurf et al., 2016, Allinen et al., 2004, Gil Del Alcazar et al., 2017). Presently, it remains unknown how these different molecular axes together contribute to DCIS evolution.

Here, as part of the NCI Human Tumor Atlas Network (HTAN) we collected and curated two of the largest DCIS cohorts to date from the Translational Breast Cancer Research Consortium (TBCRC) 038 study and the Resource of Archival Breast Tissue (RAHBT), on which to conduct comprehensive molecular analyses. We performed a multimodal integrated profile of these complementary, longitudinally sampled DCIS cohorts, in order to understand the spectrum of molecular changes in DCIS and to identify predictors of subsequent events in both tumor and stroma. We used multidimensional and multiparametric approaches to address the central conceptual themes of cancer progression, ecology and evolutionary biology, and molecular subtypes. Multiple data types were applied to create a platform for complex multi-dimensional data representation. We hypothesize that the breast precancer atlas (PCA) presented here will allow for the application of phylogenetic tools that can reconstruct the relationship between DCIS and IBC, the natural history of DCIS, and factors that underlie progression to invasive disease.

RESULTS

Study Design and Cohorts

We generated two retrospective study cohorts of patients with DCIS. Each cohort was composed of cases with DCIS who had no later events, and cases with DCIS who had a subsequent ipsilateral breast event (iBE, either DCIS or IBC) after surgical treatment. **Table 1**

summarizes the cases in both cohorts analyzed in this study. The two cohorts had complementary strengths. Samples from the TBCRC cohort were macrodissected to enrich for DCIS epithelial cells for downstream RNA and DNA profiling with SMART-3SEQ (Foley et al., 2019) and low-pass whole genome sequencing (WGS), respectively. In contrast, the RAHBT cohort was organized into a tissue microarray (TMA), with laser capture microdissection (LCM) used to precisely separate samples into epithelial and stromal components (**Figure S1A-L**). These were sequenced separately for RNA analysis. DNA analysis was performed on LCM epithelium only. Adjacent normal epithelial and stromal samples were obtained by LCM. Sequential TMA tissue sections were used for multiplexed ion beam imaging (MIBI, **Figure 1**). Patient composition differences between TBCRC and RAHBT were reflected in variable predictions of iBEs by known prognostic biomarkers, such as ER and HER2 expression (**Figure S1M-P**). **Table S1** summarizes the full RAHBT and TBCRC cohorts, and **Table S2** summarizes the assays used in this study by cohort.

Expression and genomic analyses reveal molecular differences between normal breast, DCIS, and IBC epithelium

We first examined the changes that accompany the progression from normal breast to DCIS and IBC using LCM obtained RAHBT epithelial samples. Uniform Manifold Approximation Projection (UMAP, McInnes et al., 2018) analysis of the top 500 most variably expressed genes (**Figure 2A**) illustrates the spectrum of DCIS distributed amongst IBC and normal samples. Whereas most IBCs were interspersed amongst the DCIS samples, indicating a high degree of heterogeneity in both DCIS and IBC, most normal samples clustered tightly together.

To specifically define the progression from *normal epithelium to DCIS*, we analyzed the expression differences between normal breast tissue (n=28) and primary DCIS (n=243). We identified 1660 differentially expressed (DE) genes (DESeq2, **Figure S2A**), of which many have been shown to be involved in transition to a cancer phenotype. By gene set enrichment analysis (GSEA), normal epithelium was enriched in pathways related to epithelial integrity such as biological adhesion, regulation of cell population and proliferation, anchoring- and apical junctions, tube morphogenesis, and regulation of cell differentiation (**Figure 2B**). Conversely, DCIS epithelium was enriched in pathways associated with cell growth and

neoplasia including cell cycle, oxidative phosphorylation, estrogen response, metabolism, MYC targets, mTORc1 signaling, and DNA repair (**Figure 2B**).

Next, we analyzed transcriptional differences between non-matched *primary DCIS (n=243) and IBC (n=43)* epithelial samples. We identified 170 DE genes (DESeq2), amongst which were a group of mesenchymal genes, including vimentin and collagens, that were increased in the epithelial component of IBC. Pathway analysis confirmed that the main enriched pathways in IBC were related to epithelial mesenchymal transition (EMT) and extracellular matrix (ECM, **Figure 2B**). We leveraged MIBI data to verify the observed mesenchymal gene expression in IBC epithelial samples at protein level (**Figure S2B).**

Having examined the transcriptomic differences, we compared the copy number alterations (CNAs) in DCIS to IBC to identify genomic alterations involved in invasion. We profiled the CNA landscape of 58 IBC and 406 DCIS samples across TBCRC and RAHBT cohorts using low-pass WGS. To adjust for differences in epithelial content between the two cohorts, CNA calls were corrected for extent of tumor content, *i.e.* tumor purity (see **Methods**). Overall, the CNA landscape of DCIS closely reflects IBC, in contrast to the quiet genomic landscape of normal breast tissue (**Figure 2C**). Similar CNA landscapes were observed when considering DCIS and IBC samples from each cohort individually (**Figure S2C**). IBC samples had higher genomic instability compared to DCIS, as quantified by the Proportion of the Genome copy number Altered (PGA; $\log_2$FC = 0.60; P= 1.40×10^{-2}; **Figure 2D**). Increased genomic instability in IBC *vs* DCIS was observed in both TBCRC and RAHBT cohorts (**Figure S2D**).

Finally, we examined protein expression by MIBI with a 37-plex antibody panel for markers of differentiation and oncogenic signaling, to define cellular composition and structural characteristics of normal breast, DCIS, and IBC samples. UMAP analysis demonstrated an expansion of luminal enriched clusters in the DCIS and IBC cells compared to normal breast epithelium (**Figure 2E-F**). As found in the RNA and DNA data this was accompanied by an increase in ER and HER2 activity (**Figure 2G**). Further analysis of the MIBI data is discussed in detail in the companion paper by Risom et al.

Three unique transcriptional subtypes characterize DCIS biology

Genomic and transcriptomic-based classifications of IBC (Perou et al., 2000, Curtis et al., 2012) have crystallized our understanding of the spectrum of biologic subtypes in invasive breast cancer, but it remains unclear whether these classification schemes accurately describe the spectrum of the DCIS stage. To answer this question, we applied the PAM50 classification to TBCRC DCIS samples and RAHBT DCIS epithelial samples. Luminal A was the most abundant subtype in both cohorts, followed by Basal-like, HER2-enriched, Luminal B and normal-like (**Table S3**). To assess how well DCIS aligned with the PAM50 classification, we evaluated the correlation of each DCIS tumor to the centroid of its assigned subtype and compared the correlations against those observed in 1109 IBCs from The Cancer Genome Atlas (TCGA; **Figure 3A**). The median correlation (Spearman's ρ) for basal-like DCIS samples was significantly lower than basal-like IBC samples (median ρ_{IBC} = 0.75; median ρ_{DCIS}= 0.38; $P_{Bonferroni}$ = $8.01x10^{-16}$; Wilcoxon rank sum test), as previously shown (Bergholtz et al., 2020). Significantly decreased correlation was also observed for luminal A (median ρ_{IBC} = 0.60; median ρ_{DCIS}= 0.50; $P_{Bonferroni}$ = $3.13x10^{-3}$) and normal-like subtypes (median ρ_{IBC} = 0.60; median ρ_{DCIS}= 0.49; $P_{Bonferroni}$ = $6.21x10^{-3}$). Notably, projecting the DCIS transcriptome onto two-dimensions using UMAP revealed clear deviations from the PAM50 centroids (**Figure 3B; Figure S3A**). PAM50 subtypes failed to distinguish cases with and without iBEs (**Figure S3B-C,** P_{TBCRC} = 0.274; P_{RAHBT} = 0.211; **Table S3**). Similarly, the integrative subgroups (ICs), also developed for IBC based on integration of genomic copy number and expression profiles (called here using expression data alone), failed to robustly predict recurrence in DCIS (**Figure S3D**). These data suggest that while established IBC subtypes can be identified in DCIS, they do not fit DCIS as robustly as IBC, and are not prognostic in these premalignant lesions.

The limited utility of IBC-based subtypes in DCIS presented an opportunity to identify DCIS-specific subtypes, an approach restricted until now by small sample size in previous cohorts. We discovered de novo DCIS-specific subtypes in TBCRC samples. Using non-negative matrix factorization (NMF) on all coding genes with non-zero variance, we evaluated the fit of 2-10 clusters and selected a novel three cluster solution based on optimization of silhouette width, cophenetic value, maximizing cluster number and replication in RAHBT (more on replication in RAHBT below; **Figure 3C-D; Figure S3E**). The three clusters separated well on

a UMAP of the full transcriptome (**Figure 3C**). We compared the three clusters to PAM50 subtypes as well as *ERBB2* and *ESR1* abundance which are prognostic in IBC. Cluster 1 was enriched for HER2 and basal-like PAM50 subtypes (Odds Ratio $(OR)_{HER2}$ = 18.2; P_{HER2} = $6.85x10^{-13}$; OR_{Basal} = 7.19; P_{Basal} = $1.12x10^{-8}$; Fisher's exact test). Cluster 1 had significantly higher levels of *ERBB2* (log_2FC = 2.97; FDR = $5.98x10^{-38}$; DESeq2) and lower levels of *ESR1* (log_2FC = -2.75; FDR = $3.26x10^{-36}$) compared to clusters 2 and 3 (**Figure 3D**). Clusters 2 and 3 had increased *ESR1* expression, with cluster 2 enriched for normal-like PAM50 subtype (OR = 11.2; P= $1.58x10^{-7}$; Fisher's exact test) and cluster 3 enriched for both PAM50 luminal A and B subtypes (OR_{LumA} = 10.9; P_{LumA} = $5.65x10^{-11}$; OR_{LumB} = 3.96; P_{LumB} = $2.94x10^{-4}$; Fisher's exact test).

Next, we replicated the three clusters in RAHBT epithelial samples to provide confidence that the transcriptional patterns reflect DCIS biology rather than technical or stochastic differences unique to TBCRC. First, we independently identified three clusters in RAHBT (**Figure 3E; Figure S3E-F**). In parallel, we identified the centroids of each of the three TBCRC clusters (n_{genes} = 1,350) and applied them to RAHBT. De novo clustering and TBCRC-based centroid predictions in RAHBT were highly concordant (concordance = 0.80; **Figure S3G**) and had significantly higher silhouette widths, indicating better fit, than PAM50 (P= $4.2x10^{-23}$; Wilcoxon rank sum test; **Figure S3H**). We evaluated increasing the number of clusters to six, which was the largest number of clusters before we observed a decrease in cluster fit (**Figure S3E**). Six clusters further separated clusters 1 and 3 but failed to show concordance across cohorts (**Figure S3I**). Therefore, we focused on characterizing three transcriptional DCIS subtypes. Similar to TBCRC, cluster 1 in RAHBT had significantly higher levels of *ERBB2* (log_2FC = 2.66; FDR = $9.48x10^{-32}$; DESeq2) and lower levels of *ESR1* compared to clusters 2 and 3 (log_2FC = -1.87; FDR = $4.73x10^{-15}$). Henceforth, we referred to the three clusters as ER_{low}, quiescent, and ER_{high} respectively.

Validation of DCIS subtypes identifies a metabolically quiet subtype

To characterize the three DCIS subtypes, we conducted differential abundance analysis comparing each cluster individually to the other two combined (*i.e.* one-vs-rest). The deregulated pathways in each cluster were highly concordant across cohorts, further supporting three transcriptional patterns in DCIS (P_{ERlow} = $2.33x10^{-2}$; $P_{quiescent}$ = $8.37x10^{-2}$; P_{ERhigh} = $9.20x10^{-10}$; hypergeometric test; **Figure 3F**). We observed an upregulation of

estrogen response in the ER_{high} subtype, while the ER_{low} subtype was upregulated for mTOR signaling. In support, we investigated protein expression by MIBI for a subset of these patients (n=71). The frequency of ER+ tumor cells was significantly higher in the quiescent and ER_{high} subtypes compared to ER_{low} (log_2FC = 2.73; P= $2.11x10^{-5}$; Wilcoxon rank sum test) while HER2+ tumor cells were significantly higher in the ER_{low} subtype (log_2FC = 4.88; P= $3.74x10^{-2}$; Wilcoxon rank sum test; **Figure 3G**). In general, the frequencies of ER and HER2+ tumor cells were well correlated with RNA abundance of *ESR1* and *ERBB2*, respectively (**Figure S3J-K**). *PGR* levels were similarly upregulated in quiescent and ER_{high} compared to ER_{low} ($log_2FC_{quiescent}$ = 1.01; $FDR_{quiescent}$ = $6.28x10^{-3}$; log_2FC_{ERhigh} = 1.89; FDR_{ERhigh} = $4.43x10^{-6}$; DESeq2; **Figure S3L**).

In TBCRC, the quiescent subtype was enriched for the normal-like PAM50 subtype, however, in RAHBT it was enriched for luminal A (OR = 11.8, P= $9.95x10^{-7}$; Fisher's exact test). This is likely due to the lack of stromal signal in the epithelial-enriched RAHBT RNA sequencing compared to TBCRC, which contains both stromal and epithelial contributions. Leveraging our MIBI data, the quiescent lesions were depleted for Ki67 (log_2FC = -1.46; P= $8.08x10^{-2}$; Wilcoxon rank sum test) and GLUT1 (log_2FC = -2.64; P= $8.47x10^{-3}$) positive tumor cells, compared to ER_{high} and ER_{low} tumors, suggesting quiescent lesions are less proliferative and less metabolically active (**Figure 3G-H**). In support, pyruvate kinase (PKM), phosphoglycerate kinase 1 (PGK1), isocitrate dehydrogenase 1 (IDH1) and fumarase (FH), four key enzymes in glycolysis or oxidative phosphorylation, were significantly downregulated in the quiescent subtype in TBCRC (**Figure S3M**). Similar downregulation of these metabolic enzymes was observed in RAHBT and the quiescent subtype most reflected normal breast tissue (**Figure S3N**). By contrast, ER_{low} lesions had significantly higher frequency of Ki67 positive tumor cells compared to quiescent and ER_{high} lesions (log_2FC = 1.03; P= $7.01x10^{-3}$; Wilcoxon rank sum test; **Figure 3G-H**). Increased proliferation of ER_{low} and decreased proliferation of quiescent lesions was recapitulated using a transcriptomic signature (Venet et al., 2011) in both cohorts (**Figure SO**).

Finally, we evaluated the clinical prognostic value of these three transcriptional subtypes. In the TBCRC cohort, the three subtypes were significantly associated with time to iBEs (P= $9.1x10^{-3}$; **Figure 3I**). The quiescent subtype had significantly better outcomes than ER_{high} and ER_{low} subtypes ($HR_{quiescent\ vs\ ERlow}$= 0.39; $P_{quiescent\ vs\ ERlow}$ = $4.49x10^{-3}$; $HR_{quiescent\ vs\ ERhigh}$ = 0.57;

$P_{quiescent\ vs\ ERhigh}$ = 6.65x10^{-2}; CoxPH model correcting for treatment). The ER_{low} subtype had the worst outcome, in line with previous reports of ER-negative DCIS recurring earlier than ER-positive DCIS (Meattini et al., 2017). In the RAHBT cohort, the ER_{high} subtype had better outcome than ER- similar to TBCRC (**Figure S3P**), however, the quiescent group had the worst outcome. This discrepancy with TBCRC may be due to the weaker quiescent signal in RAHBT as a result of the lack of stromal signal in the LCM epithelial samples; however, the quiescent group had the worst outcome. This discrepancy with TBCRC may be due to patient composition differences, as noted previously, (**Figure S1M-P**) and to the weaker quiescent signal in RAHBT as a result of the lack of stromal signal.

Amplifications characteristic of high-risk of relapse IBC occur in DCIS

We interrogated the genomic landscape of DCIS to identify recurrent CNAs that characterize DCIS. We identified 17 recurrent CNAs, 9 gains and 8 losses, in DCIS occurring in 17.0-52.5% of samples (FDR < 0.05; **Figure 4A**). The identification of these 17 common CNAs was not biased by depth of sequencing or cohort (**Table S4**). The most frequent alterations were gain of chromosome 1q and 17q, particularly 17q12 where the *ERBB2/HER2* oncogene is located, and loss of chromosome 16q and 17p (**Figure 4A**), confirming prior findings (Yao et al., 2006, Lesurf et al., 2016, Abba et al., 2015, Trinh et al., 2021) and notably reflecting the CNA landscape of IBC (Russnes et al., 2010, Curtis et al., 2012). Next, we investigated if the transcriptomic DCIS subtypes biased the CNA landscape. We observed a strong enrichment of amplifications of chr17q12 (*ERBB2*) in ER_{low} samples compared to ER_{high} and quiescent samples (OR = 4.43; P=8.19x10^{-12}; Fisher's exact test; **Figure 4A**). Deletion of 6q16.1 (OR = 1.67; P= 3.15x10^{-2}), 11q25 (OR = 2.46; P= 1.83x10^{-5}) and 16q23.3 (OR = 2.84; P= 3.40x10^{-7}) were enriched in the ER_{low} subtype. We observed a significant difference in PGA across the DCIS subtypes. The ER_{high} subtype had the highest PGA ($log_2FC_{ERhigh\ vs\ others}$ = 0.36; P= 1.05x10^{-3}; Wilcoxon rank sum test) while the quiescent subtype had the lowest PGA ($log_2FC_{quiescent\ vs\ others}$ = -0.50; P= 8.88x10^{-3}; **Figure 4B**). PGA was not biased by sequencing depth (**Figure S4A**). Finally, we investigated if any of the 17 recurrent CNAs were predictive of DCIS or IBC iBEs. We did not identify any robust associations between CNAs and risk of iBEs in DCIS (**Figure S4B**). PGA was not predictive of iBEs (**Figure S4C**), and no significant difference was observed in genomic instability between cases with and

without iBEs ($log_2FC_{DCIS\ iBEs\ vs\ controls}$ = 0.12; $P_{DCIS\ iBEs\ vs\ controls}$ = 0.69; $log_2FC_{IBC\ iBEs\ vs\ controls}$ = 0.26; $P_{IBC\ iBCs\ vs\ controls}$ = 0.25; **Figure 2D**).

Early patterns of alterations may provide insight into the mechanisms by which neoplastic lesions form and progress towards invasion. Similar to the transcriptome, we employed unsupervised clustering (*i.e*., NMF) to identify genomic subtypes emerging in DCIS. Unlike the transcriptomic-based subtyping and clustering, which was sensitive to the presence or absence of stromal elements (and potentially technical differences in data generation), genomic data from FFPE samples are generally robust and not subject to these factors. Accordingly, we ran NMF on CNA segments on TBCRC (macrodissected) and RAHBT (LCM epithelia only) jointly. We identified six CNA clusters ranging in size from 6-300 samples (**Figure 4C**; **Figure S4D**). The clusters identified in TBCRC and RAHBT cohorts individually were highly concordant with clusters identified when combining the cohorts (**Figure S4E-F**) and the clusters were not biased by tumor purity (**Figure S4G**). CNA clusters 1 and 4 were characterized by amplifications of *ERBB2* on chr17q12 and were enriched for HER2+ tumors, as classified by PAM50 or the ICs (IC5) (**Figure 4C-D**). Cluster 2 was enriched for ER+ tumors (**Figure 4C**). Intriguingly, despite IC subtypes poorly predicting recurrence in DCIS (**Figure S3D**), the six CNA subtypes could be attributed to the presence or absence of CNAs characteristic of the IC subtypes, namely the four high-risk of relapse ER+/HER2- subgroups (IC1,2,6,9) and the HER2-amplified (IC5) subgroup (Rueda et al., 2019)(**Figure 4D**). For example, cluster 2 was characterized by amplifications of chr8p11.23 (similar to IC6), cluster 3 by amplifications in chr11q13.3 (similar to IC2) and cluster 5 by amplifications of chr20q13.2 (IC1 & IC9; **Figure 4D**). Amplifications in cluster 4 spanned beyond *ERBB2* and included regions of chr17q23.1 (IC1) which were absent in cluster 1. Finally, cluster 6 represented a CNA quiet subgroup, characterized by the absence or diminished signal of the aforementioned CNAs. Cluster 6 was the largest subgroup (n=300), in agreement with the higher genomic stability observed in DCIS compared to IBC (**Figure 2D**). The six CNA clusters were not associated with iBEs (**Figure S4H-I**) and were only weakly associated with the DCIS subtypes, primarily enrichment of the *ERBB2*-amplified clusters 1 and 4 in the HER2-enriched subtype (**Figure S4J**). Of note, these four high-risk integrative subgroups (IC1,2,6,9) account for 25% of ER+/HER2- IBC and the majority of distant relapses (Rueda et al., 2019). Integrative subtypes are prognostic in IBC and improve the prediction of late relapse relative to clinical covariates. Understanding the clinical course of DCIS lesions

harboring these high-risk invasive features is highly relevant in refining clinically meaningful risk associated with DCIS progression.

The DCIS TME is heterogeneous and reflects distinct immune and fibroblast states

Accumulating evidence has shown that the tumor microenvironment (TME) is crucial for cancer development and progression (Hinshaw and Shevde, 2019, Gil Del Alcazar et al., 2020). We used LCM-obtained stromal samples from the RAHBT cohort to analyze the relationships and crosstalk within the TME. CIBERSORTx (CSx, see **Methods, Figure S5A-C**) and MIBI derived cell type frequencies from adjacent slides from the same DCIS samples were compared. This analysis demonstrated a strong correlation for most of the cell types by both methods (**Figure S5D**).

We performed a shared nearest neighborhood (SNN) analysis and identified four distinct DCIS-associated stromal clusters (**Figure 5A-B**). Gene ontology (GO) and KEGG pathway analyses (**Figure 5C**), together with CSx (**Figure 5D**, **Figure S5E-H**) and MIBI (**Figure 5E-F**) were used to describe the major characteristics of each cluster. Cluster 1, with 421 highly-expressed genes, showed pathways involved with ECM organization, complement and coagulation cascades, focal adhesion, and PI3K-Akt signaling and was termed the “normal-like” stromal cluster. Cluster 2 had 561 highly-expressed genes, and was characterized by pathways associated with collagen metabolism, TGFb signaling, and proteoglycans in cancer, and shared genes involved in cell-substrate and focal adhesion with the “normal like" cluster and was termed the “collagen rich” stromal cluster. By predicting frequencies of cell types present in each sample we found that this cluster had the highest fibroblast abundance and total myeloid cells, mostly associated with macrophages and myeloid dendritic cells (mDC). MIBI results showed that this cluster was enriched in collagen and fibroblast associated protein positive (FAP+,VIM+, SMA+) fibroblasts. Cluster 3, with 702 upregulated genes, was represented by mammary gland development and fatty acid metabolism, high presence of CD8 T cells assessed by CSx, and myofibroblasts by MIBI (VIM+, SMA+ fibroblasts) and termed the “desmoplastic” stromal cluster. Cluster 4 presented the largest number of upregulated genes (1244), associated with immune response and was termed the “immune dense” stromal cluster. We confirmed by both CSx and MIBI that total abundance of

immune cells was more than twice compared with the other clusters, with predominance of lymphoid over myeloid cell populations. Within this cluster, a subgroup of samples was highly enriched for B cells, whereas another subgroup showed an overall balanced immune cell type composition. This cluster correlated with PAM50 Basal and HER2 (Chi^2 test, P_{basal} = 2.66x10-7, P_{HER2} = 1.23x10-5) subtypes and was enriched in the ER_{low} RNA cluster (Chi^2 test, P= 4.23x10-8, **Figure 5D**, **Figure S5I-J**). **Figure 5E** shows a representative MIBI image of each cluster, where a strong correlation with fibroblast states and immune cell density is observed.

Stromal subtypes and stromal cell states had only a modest effect on outcome. Although not significant, cluster 2, characterized by fibroblasts and CD8 T cells, had a lower risk of iBEs compared to the other three clusters (**Figure S5K-L**). Furthermore, by univariate CoxPH analysis, we found that monocyte abundance was associated with an increased risk of DCIS recurrence in RAHBT (adj. P= 0.012, **Table S5**).

We further analyzed the cluster-associated DCIS epithelial samples to investigate the relationship between epithelial and stromal features. We found no significant difference in proliferation index (Venet et al., 2011), EMT score (Mak et al., 2016), or PGA between the clusters (**Figure 5M-O**). By differential gene expression analysis (DESeq2, *one-vs-rest*), DCIS epithelium associated with “immune dense” stroma had 300 DE genes (**Figure S5P**) whereas the other groups were almost indistinguishable, with 11, 29, and 2 DE genes. GO analysis of the 300 DE genes revealed upregulated immune pathways (neutrophil activation, cell chemotaxis, antigen processing and presentation), ribose phosphate metabolic process, and apoptosis, which could explain the augmented immune cell recruitment in the stroma (**Figure S5Q**).

Last, we identified the centroids of each stromal cluster and applied them to non-matched normal (n=10) and IBC (n=30) stromal samples (**Figure 5G**). We found that all normal and two IBC samples were predicted as ‘normal-like’ stroma, whereas the remaining IBC samples were evenly distributed across the collagen rich, desmoplastic and immune dense clusters, suggesting that IBC and DCIS TMEs share stromal patterns. *De novo* clustering of normal and IBC stromal samples was highly correlated with predicted clusters obtained by the centroids method (**Figure S5R).** Interestingly, transcriptomic (DESeq2, one-vs-rest) and

GO analysis of normal-like stroma vs true normal stroma showed that normal-like stroma differs on cortisol and steroid signaling, response to mechanical stimulus and epithelial cell proliferation (**Figure S5S**).

Together, these results reflect different DCIS-associated stromal reactions, mostly characterized by different fibroblast states and immune cell density. These findings are in agreement with those observed by MIBI in RAHBT samples (Risom *et al*.), where loss of normal, resting fibroblast was observed in the progression from normal to DCIS and IBC, coupled with an increase in CAFs.

Oncogenic pathways characterize poor outcome groups

The TBCRC and RAHBT cohorts were designed for outcome analysis, with inclusion of both patients with subsequent iBEs, and patients that did not have any events during long term follow up. We used TBCRC samples to identify gene expression patterns that correlate with outcome. To identify differentially expressed genes in cases with vs without subsequent iBEs, we analyzed primary DCIS with iBEs within 5 years (n=72) vs. the remaining samples (n=144) from the TBCRC cohort to avoid including non-clonal events that might be more common in later years. Suspecting that the resulting 812 DE genes (DESeq2) represent multiple routes to subsequent iBEs, we leveraged NMF to identify paths to progression. Maximizing the silhouette value and number of clusters, we identified 4 clusters in TBCRC (**Figure 6A, Figure S6A-B**). Here, two clusters were associated with low rates of iBEs while two clusters, one ER+ and one ER-, were associated with higher iBE rates (**Figure 6A-B**). Cluster 1 was predominantly ER-, HER2+, and enriched in comedo necrosis, DCIS grade 3 tumors, the ER_{low} de novo RNA cluster, and Basal-like and HER2 PAM50 subtypes. Cluster 2 was enriched in the normal-like PAM50 subtype, and the Quiescent RNA cluster. Clusters 3 and 4 were both highly enriched in the ER_{high} RNA cluster and luminal PAM50 subtypes (all P<0.001, **Figure 6A, Figure S6C**). Survival analysis showed that cases in cluster 2 and 4 had significantly longer time to iBEs compared to cluster 1 and 3 (**Figure 6B**), and this difference was also significant after adjusting for treatment (cluster 2: HR (95% CI): 0.25 (0.13 – 0.47), P<0.001. Cluster 4: HR (95% CI): 0.51 (0.32 – 0.84), P=0.007, **Table S6A**).

The same survival trends were observed when considering subsequent DCIS or IBC events individually (P_{DCIS} = 0.00038; P_{IBC} = 0.013; **Figure S6D-E, Table S6B-E**).

To elucidate the molecular basis of the four clusters, we performed gene expression analysis (DESeq2) followed by GSEA of each cluster vs. the rest. Cluster 1 showed enriched mTORc1 signaling, MYC target and inflammatory response, together with reduced estrogen response, consistent with it being dominated by ER- samples. Cluster 2 also showed modest reduced estrogen response, in addition to enrichment of genes involved in hypoxia, IL2-STAT5 signaling, KRAS signaling, and inflammatory response. Cluster 2 was also the only group with apical junction enrichment. Cluster 3 showed increased oxidative phosphorylation, MYC targets, and estrogen response, consistent with the ER status of this group, whereas cluster 4 showed increased estrogen response, but reduced glycolysis, hypoxic-, and inflammatory response (**Figure 6C**).

Differences in the transcriptional profiling strategies and cohort composition between TBCRC and RAHBT make it challenging to port signatures across the two cohorts. The RAHBT epithelial and stromal samples analyzed separately do not represent the entire tumor biology and are difficult to informatically recombine (see **Discussion**). Therefore, to further explore the prognostic value of the 812 genes identified in TBCRC, we performed NMF clustering of the RAHBT epithelial samples using this gene signature and compared the resulting clusters to those discovered in TBCRC. We identified three clusters in RAHBT (**Figure S6F-H**) with distinct outcome profiles (**Figure 6D, Figure S6I)**. Comparison of the NMF weights for each cluster in TBCRC *vs* RAHBT showed strong correlation between TBCRC and RAHBT clusters 1, 2, and 3, respectively (all: Pearson's R=0.77, P<0.0001, **Figure 6E**), indicating that the respective clusters identified in each cohort were driven by the same subset of genes from the 812 gene set. RAHBT cluster 1 samples had mixed ER and HER2 status, were largely classified as the Quiescent RNA cluster and Luminal A PAM50 subtype, and were enriched for desmoplastic stroma (**Figure S6J**). Cluster 2 was enriched for ER-, HER2+ samples, DCIS grade 3, comedo necrosis, the ER_{low} RNA cluster, the Basal and HER2 PAM50 subtypes, and the immune dense stromal cluster. Cluster 3 was largely ER+ with mixed HER2 status, enriched for the ER_{high} and Luminal B subtypes (**Figure S6J**). Moreover, patients in cluster 3 had a significantly higher hazard ratio compared to cluster 2 (HR (95%

CI): 2.4 (1.13 - 4.9), P=0.023, **Figure 6D**), although this was not significant after adjusting for treatment (P=0.125, **Table S6F**). Of note, TBCRC cluster 4, which was characterized by high ER signaling, but low hypoxic and inflammatory response, and low IL2-STAT5 signaling, was not identified in the RAHBT cohort, probably due to the different sample composition and technical differences between these cohorts.

Next, we performed outcome analysis specifically for DCIS iBEs in RAHBT. We observed fewer DCIS iBEs in cluster 1 than cluster 2, contrary to observations in the full cohort (P=0.065, **Figure S6K**), although the difference was not statistically significant (**Table S6G-H**). For the analysis in RAHBT IBC iBEs, the three groups overall displayed the same trend as in the full cohort (P=0.064, **Figure S6L**), with cluster 3 showing significantly higher hazard ratio than cluster 2 (HR (95% CI): 3.3 (1.2 - 9.5), P=0.024, **Table S6I**), which was borderline significant after adjusting for treatment (P=0.058, **Table S6J**).

HER2 has been suggested as a biomarker for aggressive DCIS (Mustafa et al., 2017, Di Cesare et al., 2017). Here, we took advantage of the multiomic data available for the RAHBT cohort to investigate *ERBB2* (17q12) amplification by WGS, and *ERBB2*/HER2 expression by RNA-seq and MIBI, in the context of the NMF clusters (based on the 812 gene set). By MIBI HER2all, which corresponds to the clinical definition of HER2 positivity (cutoff >0.2, see Risom *et al*.), we observed higher HER2 expression in cluster 2 vs cluster 1 (P=0.033) and cluster 3 (P=0.024, Wilcoxon rank sum test, **Figure 6F**). Moreover, the expression observed by MIBI correlated with RNA expression (Pearson's R=0.63, P=5.2e-11, **Figure 6F**). Looking at HER2 intense expression by MIBI (cutoff >0.7), we observed a striking difference between cluster 2 and both cluster 1 and 3 (both P=0.024, Wilcoxon rank sum test, **Figure 6G**). The HER2 intense protein staining and RNA *ERBB2* expression were highly correlated (Pearson's R=0.79, P<2.2e-16, **Figure 6G**). Thus, we observed moderate HER2 expression in cluster 3, characterized as the poor outcome group, whereas *ERBB2* amplification and concordant high HER2 expression were found almost exclusively in cluster 2, a group characterized by the most favorable outcome. The observation of high HER2 expression not being correlated with poor outcome is in line with a previous study, where HER2 overexpression was found inversely correlated with the progression of DCIS to IBC (Lin et al., 2019). For further analysis of HER2 protein expression in RAHBT, see the companion paper by Risom *et al*.

Taken together, we here identified 812 genes that when used for unsupervised clustering characterized groups with similar underlying biology in both cohorts. Poor outcome groups in both cohorts exhibited increased ER and MYC signaling and oxidative phosphorylation, suggesting these pathways are important for DCIS recurrence and progression.

Finally, our interrogation of the DCIS Atlas suggests multiple routes to iBEs with varying contributions from genomic, transcriptomic and microenvironmental perturbations. To formally assess the contributions of each, we trained an elastic net model to predict time to iBEs in 80% of the RAHBT cohort and tested its performance in a held-out 20% testing cohort. The RAHBT cohort was uniquely microdissected, facilitating profiling of the TME that could not be recapitulated in TBCRC. We trained three independent models reflecting the genomic, transcriptomic and TME contribution to iBEs and assessed their performance (**Figure S6L**; see **Methods**). To ensure robust estimates of performance, we trained each model 100 times and averaged the C-index on the test cohort. The transcriptomic model (RNA; C-index = 0.57; 95% CI = 0.50-0.63; 100 iterations) performed similar to the genomic model (CNA; C-index = 0.56; 95% CI = 0.53-0.58) followed by the TME model (C-index = 0.51; 95% CI = 0.40-0.55; **Figure 6H**). Next, we evaluated the performance of a multivariate model (MV) integrating transcriptomic, genomic and microenvironmental features that were nominally associated with iBEs in the training cohort (P <0.1; CoxPH model; $n_{RNA} = 7$; $n_{CNA} = 2$; $n_{TME} = 4$). The MV model performed the best across all four models (C-index = 0.69; 95% CI = 0.68-0.71; **Figure 6H**). These data suggest genomic, transcriptomic and microenvironmental alterations contribute to iBEs. Next, we evaluated the contribution of each of the feature to iBEs in the multivariate model (**Figure 6I**). The TME subtypes were the most informative feature followed by CD4 T cell abundance in the stroma and deletions of 6q16.1, both of which were associated with longer time to an iBE. Conversely, *EDN2* (endothelin 2) abundance, previously implicated in breast cancer progression (Grimshaw et al., 2004), and increased fibroblast signal in the tumor epithelial, an indication of EMT, were predictive of shorter time to an iBE. These data point to a fundamental interplay between the transcriptome, genome, and microenvironment during DCIS recurrence or progression to IBC.

DISCUSSION

The Aims of the HTAN Breast Pre-Cancer Atlas project are to 1) develop a resource of multi-modal spatially resolved data from breast pre-invasive samples that will facilitate discoveries by the scientific community regarding the natural history of DCIS and predictors of progression to life-threatening IBC; and 2) populate that platform with data from retrospective and longitudinal (watchful waiting) cohorts of patients with DCIS and demonstrate its use to construct an atlas to test novel predictors of progression. Generating an atlas of DCIS is similar to the effort of TCGA for IBC. However there are important differences and challenges in DCIS. First, obtaining DCIS tissue samples is considerably more challenging. In IBC, the tumor is evident by gross exam, and can be easily obtained as fresh, fresh frozen, or archival material. This is not the case for pre-invasive lesions. DCIS can sometimes be recognized radiographically but is only precisely detailed by pathologic examination. An additional complication in the study of DCIS is the ambiguous nature of the process. In IBC, the transition from an intraepithelial neoplasia to an invasive neoplasia is definitional. For DCIS, such a clear-cut definition does not exist. DCIS is broadly defined by significant cytologic and architectural changes compared to normal breast architecture by a growth of neoplastic cells in the inter-epithelial compartment.

In the current study, we examined two retrospective study cohorts comprised of 481 DCIS patients, who either had no later events, or with a subsequent ipsilateral breast event. We generated RNA gene expression profiling, DNA light pass WGS, and multiplex immunohistochemistry. Using transcriptomic data from LCM samples, we found specific gene pathways altered in the transition from normal to DCIS, and DCIS to IBC. Invasion or penetration of the basement membrane and loss of the myoepithelial layer are the key histologic features that distinguish IBC from DCIS, but the molecular basis for this has yet to be discovered. While there are systematic differences in gene expression between DCIS and invasive cancer, the most notable observed here, by both RNA-seq and MIBI, was an increased expression of ECM associated genes in IBC epithelium. This finding could be due to EMT occurring in IBC cells. Alternatively, it could be related to basal lamina synthesized by only the basolaterally located myoepithelial cells in DCIS, whereas invasive epithelial cells may make these proteins, in effect recapitulating the myoepithelial phenotype, regardless of where they are positioned.

IBC has been genomically profiled with several approaches, including the PAM50 and IC classification schemes. While DCIS and IBC are part of the same neoplastic process, there are differences in the TME, evolutionary age, and inter-observer variability in diagnostic labeling at different stages of progression. This suggested that a DCIS-specific classifier would correlate better with the biologic and clinical features of DCIS. Our analysis revealed clear deviations from the PAM50 and IC subtypes, supporting that these canonical IBC classification schemes are not apt for DCIS characterization. Previous studies have also identified problems with PAM50 in its applicability to DCIS (Bergholtz et al., 2020, Swanson et al., 2019). This was especially prominent for the basal-like subtype, which showed less 'basalness' compared to basal-like IBC, as we also observed here. With two large DCIS cohorts, we were optimally positioned to create the first de novo classification scheme for DCIS. We identified three transcriptomic subgroups of DCIS, characterized by ER signaling, proliferation and metabolism, that represent the fundamental genomic organization at this early stage of breast neoplasia. We found that this classification of DCIS more accurately captures the spectrum of DCIS biology than classification schemes derived from IBC, and can be robustly applied across cohorts and transcriptome protocols. The three subgroups identified here may represent the earliest variation in neoplasia transcriptome and may be applicable to earlier stages of neoplasia, such as hyperplasias.

There are several possible explanations for why traditional IBC classifiers do not perform well on the DCIS cohorts described here. First of all, HER2 expression is more common at the DCIS stage than at the IBC stage (Allred et al., 1992), possibly because of its ability to inhibit anoikis (Whelan et al., 2013, Gupta et al., 2019). It is plausible that the more common HER2 expression in DCIS leads to a different transcriptomic distribution compared to IBCs. Many ER- DCIS express HER2 without amplification, in contrast to IBC, where the HER2 specific (amplified) subtype is clearer. Moreover, the cells in DCIS are confined to the epithelial compartment and interact with myoepithelial cells and the basement membrane, and thus are presumably restricted by rules of differentiation that govern normal epithelial cells. This regulation could constrain the transcriptomic variability of the neoplastic cells and in turn the resulting possible subtypes. Finally, the evolutionary age of the neoplasm may influence differences in classification of DCIS and IBC. By comparing WGS data from primary DCIS and IBCs, we found that the same constellation of copy number changes was present in both, consistent with previous studies (Ma et al., 2003, Vincent-Salomon et al., 2008b, Hwang et

al., 2004). While DCIS had fewer genomic alterations than IBC, and a larger group of DCIS were classified as genomically quiescent tumors, recurrent genomic events that drive the integrative subtype classification (IC) scheme in IBC were evident as early as the DCIS stage.

While the IBC microenvironment has been extensively studied, the pure DCIS microenvironment remains less understood. Due to a largely intact basement membrane and myoepithelial cell layer, most DCIS tumor cells are not as exposed to the immune environment as IBC cells (Gil Del Alcazar et al., 2017). Studies have shown that the DCIS TME is characterized by higher immune infiltration than normal breast (Hussein and Hassan, 2006, Gil Del Alcazar et al., 2020), and both tumor infiltrating lymphocytes and macrophage abundance have been associated with high grade DCIS and necrosis (Hendry et al., 2017, Campbell et al., 2017). Conversely, the transition from DCIS to IBC is marked by a switch to a less active tumor immune environment, indicating this transition as a critical step in tumor progression for immune escape (Gil Del Alcazar et al., 2017).

A unique aspect of our study is the separate profiling of stroma and epithelial components through CSx analysis of LCM-derived stromal gene expression coupled with the in situ MIBI technique. We identified four stromal subtypes characterized by distinct gene pathways, stromal-, and immune cell composition. Moreover, specific stromal patterns were correlated with epithelial expression patterns, and particularly HER2+ and ER- DCIS were associated with a stronger inflammatory response. This could be due to coamplification of *ERBB2* (HER2) and a cluster of chemokine encoding genes on the 17q12 chromosomal region, as suggested by others (Gil Del Alcazar et al., 2017).

Fibroblasts are predominant components of breast tissue architecture. Resting fibroblast, abundant in normal breast, can be activated by tissue damage and immune cell response to participate in healing processes. Activated fibroblasts, or myofibroblasts, can induce epithelial cell proliferation, synthesize ECM, and produce cytokines and chemokines (Kalluri, 2016, Houthuijzen and Jonkers, 2018). CAFs constitute a heterogeneous group of activated fibroblasts categorized by expression of smooth muscle actin (αSMA), fibroblast surface protein (FSP1) and fibroblast-activated protein (FAP), amongst other markers (Calon et al., 2014). It is thought that CAFs can support tumor cell survival, dissemination, angiogenesis,

immune suppression, and therapy resistance (Houthuijzen and Jonkers, 2018). We observed a progressive increase in CAFs and decrease in normal fibroblasts from normal tissue through DCIS to IBC, based on both MIBI and RNA analysis. This and other protein data are presented in the companion paper by Risom et al, with systematic analyses of how different phenotypic and structural properties of the DCIS TME change with progression to IBC.

The current lack of specific biomarkers to distinguish low risk DCIS from those with a high risk of progression leads to significant overtreatment (Groen et al., 2017). Given the inclusion of cases with and without iBEs in both our cohorts, we ascertained whether our data could be used to predict outcome. An 812 rapid recurrence (≤5 years of primary DCIS diagnosis) gene set identified in TBCRC was used to classify both cohorts, which defined a poor outcome group (high risk of subsequent iBEs) characterized by activation of specific pathways including MYC targets, increased oxidative phosphorylation, and estrogen response, in both cohorts. Other features, such as DCIS grade and necrosis, were not enriched amongst the clusters matched across the cohorts, indicating that the 812 gene set can potentially identify cases that do not share classical histopathological traits associated with high risk of recurrence. However, further evaluation of this gene set in an independent cohort is warranted. While the transcriptome was more prognostic than the CNA landscape and the tumor microenvironment, a model integrating features from all three proved superior at predicting iBEs. These data point to an important interplay between the transcriptome, genome and tumor microenvironment in DCIS progression that is missed when assessing each in isolation. Another consideration is the variable risk of mortality for iBEs. iBEs involving indolent IBCs are less clinically consequential than iBEs involving aggressive IBCs. We found amplicons associated with late relapse IBC present in DCIS. While these were not prognostic for iBEs, later IBC events from these DCIS are likely to be more significant than other IBC events, with clear clinical implications for risk stratification and treatment approaches for DCIS.

CONCLUSION

The studies presented here provide new insight into potential etiologies of DCIS biology that we hope will guide development of future diagnostics and serve as a template for conducting

similar analyses of preinvasive cancers. The Breast Pre-Cancer Atlas is intended to address use case scenarios for future research. This includes fitting models to estimate the main evolutionary and ecological parameters of breast neoplastic progression, reconstructing the natural history of DCIS, developing a classification system for the evolvability and ecology of DCIS for risk stratification, and registering these data with clinical images.

LIMITATIONS OF THE STUDY

The Breast Pre-Cancer Atlas presented here provides a foundational advancement in the study of precancerous lesions and will be a valuable resource for years to come. Its utility comes, in part, from the inclusion of two independent and large-scale cohorts that have important and distinct differences. For example, the two cohorts represent subjects from different geographical sites and median years of diagnosis (RAHBT: 2002; TBCRC: 2008) leading to differences in follow-up (RAHBT: 100; TBCRC: 74 months) and time to recurrence (RAHBT: 62; TBCRC: 48 months). Tumor grade was lower in the RAHBT cohort which may be due to systematic differences in DCIS grading over time (Allred, 2010). There were no significant differences in age at diagnosis or treatment across cohorts. However, future observations on a cohort of patients with DCIS who are undergoing watchful waiting would provide outcome results that may be more aligned with emerging treatment strategies of DCIS. Technical differences in the handling of the two cohorts are extremely relevant in the way our data are analyzed and presented. Thin sections from TMAs (RAHBT) or whole-slides (TBCRC) from archival FFPE blocks were used as source material for all assays. Importantly, in RAHBT, after pathology review, separate areas containing tumor epithelia and stroma were subjected to LCM providing pure populations. This strategy allowed us to identify a series of stromal subtypes that could not be inferred from bulk tumor samples. Further, the purity of normal, DCIS, and invasive tumor cells yielded clean comparisons of these three epithelial states. In contrast, the TBCRC cohort comprises pathology guided macrodissection of distinct areas containing DCIS and has lower tumor cell purity than RAHBT. The variable levels of epithelia and stroma contained in the TBCRC samples is more analogous to commonly used IBC cohorts such as TCGA (https://www.cancer.gov/tcga.) and METABRIC (Curtis et al., 2012), allowing more direct comparisons. Further, the various IBC expression based classification systems were established and validated on bulk tumor samples.

Therefore, we consider the systematic technical differences between the cohorts a significant strength of the study as it allows for a broader range of downstream analyses.

ACKNOWLEDGMENTS

R01 CA185138-01 (ESH); U2C CA-17-035 Pre-Cancer Atlas (PCA) Research Centers (ESH, RBW, CM, KP, GAC); DOD BC132057 (ESH, CM); BCRF 19-074 (ESH); BCRF 19-028 (GAC) PRECISION CRUK Grand Challenge (JW); R01CA193694 (RBW, GAC), BCRF PPI-18-006 (RBW). SHS was supported by the Lundbeck Foundation (R288-2018-35) and the Danish Cancer Society (R229-A13616). KEH was supported by a CIHR Banting Postdoctoral Fellowship. TBCRC 038 was conducted by the TBCRC, which receives major funding support from The Breast Cancer Research Foundation and Susan G. Komen.

AUTHOR CONTRIBUTIONS

Conceptualization: ESH, RBW, CM, GAC, JRM, CC, and KP. Investigation: SHS, BRG, KEH, JAS, TR, SVe, AK, LC, DJV, KD, DYD, DZ, JL, SJ, SS, AH, and RBW. Resources: LAS, TH, BH, FC, KG, MK, SW, AD, TK, PM, JN, JL, JTs, AMS, AT, GG, CZ, MM, JW, SXZ, and SVa, Data curation: LK and JTa; Writing – Original Draft: SHS, BRG, KEH, JRM, ESH, and RBW. Writing – Review & Editing: All co-authors. Funding Acquisition: ESH, RBW, GAC, and CM. Project Administration: RB, LA, and CS. Supervision: CC, RT, RMA, KO, KP, CM, JRM, GAC, ESH, and RBW.

DECLARATION OF INTERESTS

CC serves on the Scientific Advisory Board and/or as a consultant for GRAIL, Deepcell, Ravel, Viosera, NanoString, Genetech and holds equity in GRAIL, Deepcell, Ravel.

KP serves on the Scientific Advisory Board of Acrivon Therapeutics, Vividion Therapeutics, and Scorpion Therapeutics, holds equity in Scorpion Therapeutics and Vividion Therapeutics, is a consultant to Aria Pharmaceuticals, and received honorarium from Astra-Zeneca.

RMA is an inventor on patent US20150287578A, and is a board member and shareholder in IonPath Inc. TR and RBW have consulted for IonPath Inc.

Figure Legends

Figure 1. Method Outline

Two retrospective study cohorts (RAHBT, TBCRC) were generated, consisting of DCIS patients with either a subsequent ipsilateral breast event (iBE) or no later events after surgical treatment. TBCRC samples were macrodissected for downstream RNA and DNA analyses. RAHBT samples were organized into a TMA and serial sections were made for RNA, DNA and protein (MIBI) analysis. For RNA and DNA sequencing, TMA cores were laser capture microdissected after tissue masking by a pathologist, to ensure pure epithelial and stromal components.

Figure 2. Comparison of normal breast, DCIS, and IBC by RNA-seq, WGS and MIBI

A) UMAP projection of the top 500 most variably expressed genes in normal (green), DCIS (red), and IBC (blue) epithelial samples. **B**) Enriched pathways (GSEA Hallmarks and Gene Ontology) in Normal and DCIS epithelial tissue (from DESeq2 analysis, normal vs. DCIS), and IBC epithelium (from DESeq2, IBC vs DCIS). Size of the dot and color represents the magnitude and direction of pathway deregulation, *i.e.* blue indicates the pathway is downregulated while red indicates the pathway is upregulated. Background shading indicates the false discovery rate. **C)** CNAs in Normal (top), DCIS (middle), and IBC (bottom). **D)** CNA burden in IBC and DCIS by outcome groups. **E)** UMAP plot of epithelial cells in normal, DCIS, and IBC tissues based on protein expression of ECAD, PanKRT, KRT7, KRT5, VIM, ER, AR, HER2, Ki67, SMA, HH3 as measured by MIBI. Blue: DCIS. Yellow: IBC. Pink: Normal. **F)** UMAP plot from *E* showing cells specifically from normal (left), IBC (middle), or DCIS (right) tissues. **G)** Cells are overlaid with their expression of ER (left), or HER2 (right) as measured by MIBI.

Figure 3. *De novo* transcriptomic DCIS subtypes

A) Invasive breast cancer intrinsic subtypes do not fit DCIS. Boxplot shows Spearman ρ of DCIS and IBC samples with PAM50 centroids. Dots are colored by PAM50 subtype and the covariate along the top indicates if the sample is IBC or DCIS. Boxplot represents median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range. **B)** UMAP projection of

DCIS transcriptome (TBCRC) colored by PAM50 subtype. Large circles represent the PAM50 subtype centroids. **C)** UMAP projection of DCIS transcriptome (TBCRC) colored by *de novo* DCIS subtypes. **D)** Unsupervised clustering of DCIS transcriptomes identifies three subtypes: ER-, quiescent and ER+. Heatmap depicts RNA abundance of 90 informative genes, y-axis, contributing to the three subtypes in TBCRC samples, x-axis. Barplot along the top indicates the proportion of PAM50 subtypes within each cluster. Covariates indicate integrative and PAM50 subtypes, along with *ERBB2* and *ESR1* mRNA abundance for each sample. **E)** Heatmap of DCIS subtypes in RAHBT. **F)** Pathways, y-axis, deregulated in each cluster are highly concordant in TBCRC and RAHBT. Size of the dot and color represents the magnitude and direction of pathway deregulation, *i.e.* blue indicates the pathway is downregulated while red indicates the pathway is upregulated. Background shading indicates the false discovery rate. Covariate along the top indicates the DCIS subtype and cohort. **G)** The ER- subtype is associated with more HER2+ tumor cells, as determined by MIBI, while the ER+ and quiescent clusters are associated with more ER+ tumor cells. The quiescent cluster has a low frequency GLUT1 and Ki67 positive cells. Dot color indicates *ERBB2* genomic amplification level. **H)** Representative MIBI images of the three subtypes in (**G**). White = Nuc; Blue = PanKRT; Yellow = SMA; Pink = GLUT1; Cyan = HER2; Green = ER; Red = Ki67 **I)** Kaplan-Meier plot of time to DCIS or IBC recurrence in three DCIS subtypes. CoxPH model was corrected for treatment.

Figure 4. Characteristic invasive breast cancer CNAs present in DCIS
A) Seventeen cytobands are significantly recurrently altered in DCIS. Heatmap shows $\log_2$ copy number for each of the recurrent CNAs, y-axis, in each sample, x-axis. Samples are grouped by DCIS subtype as indicated by the covariate along the top. The middle barplot shows the proportion of samples with each CNA. Grey and black represent increasing amplitudes. Finally, the barplot on the right shows the FDR from a Kruskal-Wallis test of each CNA with the three DCIS subtypes. The vertical line indicates FDR = 0.05. Four CNAs were significantly associated with the DCIS subtypes and the bar is colored by which subtype they were enriched in. **B)** The quiescent subtype had the lowest PGA. P-value from Kruskal-Wallis test. Boxplot represents median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range. **C)** Unsupervised clustering of CNA landscape identifies six clusters. Heatmap depicts $\log_2$ copy number of genomic segments, y-axis, in TBCRC and RAHBT samples, x-axis. Barplot along the top indicates the proportion of PAM50 subtypes within each cluster.

Covariates indicate integrative and PAM50 subtypes, along with *ERBB2* and *ESR1* positivity for each sample. Covariate along the right shows the chromosome of each segment. **D)** Boxplot shows $\log_2$ copy number, y-axis, across the six clusters, x-axis. Selected amplifications characterize each of the six clusters and are also characteristic of the invasive breast cancer integrative clusters, as indicated in the header of each boxplot.

Figure 5. Analysis of the tumor microenvironment in RAHBT cohort

A) UMAP projection of DCIS stromal transcriptome colored by the four identified clusters. **B)** Heatmap of the top 20 up-regulated genes for each stromal cluster. **C)** GO and KEGG pathway analysis of up-regulated genes in each cluster vs the rest. **D)** Deconvolution analysis by CSx of epithelial and stromal LCM samples grouped by stromal clusters shows different immune cell and fibroblast abundance in DCIS stromal clusters. **E)** Representative MIBI images of stromal clusters reflecting different fibroblast states and total immune density. Top left: normal-like. Top right: Collagen rich (FAP+). Bottom left: Desmoplastic (SMA+). Bottom right: Immune dense (CD45 high). H3, histone 3; VIM, vimentin; panCK, pan cytokeratin; SMA, smooth muscle actin; FAP, fibroblast activated protein. **F)** MIBI-estimated cell density within stromal clusters supports CSx findings (total n_{MIBI} = 59, n_{CSx} = 193). **G)** Predicted stromal clusters for normal breast and IBC based on centroid identification of DCIS stromal clusters, represented by UMAP along with non-matched DCIS stromal samples.

Figure 6. Outcome analysis

A) Heatmap of 812 genes and NMF clusters (k=4) in TBCRC. Top bars show ER and HER2 status, PAM50 and novel DCIS cluster classifications, DCIS grade, and necrosis. **B)** Kaplan-Meier curve (top) and forest plot of HR and 95% CIs (bottom, from CoxPH analysis) for TBCRC four cluster solution. **C)** Pathway analysis (GSEA Hallmark) of DE genes in each cluster vs. the rest. Size of the dot and color represents the magnitude and direction of pathway deregulation, (blue: downregulated; red: upregulated). Background shading indicates the false discovery rate. **D)** Forest plot for the RAHBT 3 cluster solution, HR and CI from CoxPH analysis. **E)** Dotplot of NMF cluster weights from TBCRC vs RAHBT, showing high correlation between cluster 1 (top), cluster2 (middle) and cluster 3 (bottom) in each cohort. **F)** HER2 expression by RNA-seq, WGS and MIBI in RAHBT. Box plots: HER2 all frequency by MIBI in NMF clusters (from 812 gene set). Individual P-values from Wilcoxon

rank sum test, overall P-value from Kruskal-Wallis test. Scatter plot: HER2 expression by RNA-seq (x-axis) vs MIBI HER2 all) colored by copy number status. P-values and R from Pearson's correlation. Image shows example of moderate HER2 (cyan) by MIBI. Red = VIM, yellow=SMA. **G)** HER2 expression by RNA-seq, WGS and MIBI in RAHBT. Box plots: HER2 intense frequency by MIBI in NMF clusters (from 812 gene set). Individual P-values from Wilcoxon rank sum test, overall P-value from Kruskal-Wallis test. Scatter plot: HER2 expression by RNA-seq (x-axis) vs MIBI HER2 (all) colored by copy number status. P-values and R from Pearson's correlation. Image shows example of intense HER2 staining (cyan) by MIBI. Red = VIM, yellow=SMA. **H)** Multivariate model incorporating RNA, CNA and TME features outperforms the individual models. Barplot shows C-index of each model. **I)** Averaged $\boldsymbol{\beta}$ (x-axis) from 100 trained elastic-net multivariate models for features (y-axis) included in at least half of the trained models. Covariates indicate the type of feature and the tissue in which the feature was measured.

Supplementary Figure Legends

Figure S1: LCM dissection of DCIS and IBC epithelium and associated stroma in RAHBT, and outcome analysis by ER and HER2 status in RAHBT and TBCRC, supplemental to Figure 1 and Table 1.

A) Marked DCIS epithelium (blue) prior to dissection. **B**) Dissected DCIS epithelium on cap. **C**) Remaining tissue on slide after LCM dissection of DCIS epithelium. **D**) Marked stroma (yellow) adjacent to dissected DCIS epithelium (blue, panel A-C) prior to dissection. **E**) Dissected stroma on cap. **F**) Remaining tissue on slide after LCM dissection of DCIS epithelium and adjacent stroma. **G**) Marked IBC epithelium (blue) prior to dissection. **H**) Dissected IBC epithelium on cap. **I**) Remaining tissue on slide after LCM dissection of IBC epithelium. **J**) Marked stroma (red) adjacent to dissected IBC epithelium (panel G-I) prior to dissection. **K**) Dissected IBC-associated stroma on cap. **L**) Remaining tissue on slide after LCM dissection of IBC epithelium and adjacent stroma. All images were taken at 2X magnification. **M-N**) Kaplan-Meier plot of time to iBE stratified by ER (*ESR1*) expression in

TBCRC (**M**) and RAHBT (**N**). **O-P**) Kaplan-Meier plot of time to progression stratified by HER2 (*ERBB2*) expression in TBCRC (**O**) and RAHBT (**P**).

Figure S2: Comparison of normal breast, DCIS, and IBC by RNA-seq, WGS and MIBI, supplemental to Figure 2.

A) Volcano plot showing DE genes in DCIS vs normal breast epithelium. Log2FC>0 mark genes up in DCIS vs. normal, and vice versa. **B**) Image showing IBC sample with high mesenchymal gene expression by MIBI. White = Nuc; Cyan = PanKRT; Yellow = SMA; Green = COLI; Red = VIM. **C**) Average CNA landscape for normal (top), DCIS (middle) and IBC (bottom). Y-axis shows the proportion of samples with a gain (red) or loss (blue). **D**) Density plot of proportion of the genome copy number altered (PGA) in IBC samples (orange), DCIS samples with IBC recurrence (green), DCIS samples with DCIS recurrence (purple) or DCIS samples with no recurrence (yellow).

Figure S3: Characterization of RNA Subtypes, supplemental to Figure 3.

A) UMAP projection of DCIS transcriptome (RAHBT) colored by PAM50 subtype. Large circles represent the PAM50 subtype centroids. **B-C**) PAM50 do not robustly predict progression in DCIS. Kaplan-Meier plots of time to progression in TBCRC (**B**) and RAHBT (**C**). **D**) IC10 subtypes do not robustly predict progression in DCIS. Kaplan-Meier plot of time to progression in TBCRC. **E**) NMF diagnostics supported three clusters in DCIS. Scatterplots show cophenetic and silhouette values with increasing numbers of clusters in TBCRC and RAHBT. **F**) UMAP projection of DCIS transcriptome (RAHBT) colored by *de novo* DCIS subtypes. **G**) Three subtypes are highly concordant across cohorts. Mosaic plot shows concordance of de novo clustering in RAHBT *vs* clusters determined from centroids identified in TBCRC. Blue indicates an enrichment while red indicates a depletion. **H**) De novo subtypes fit DCIS subtypes better than PAM50. Boxplot shows silhouette widths of PAM50 and DCIS subtypes in RAHBT. Boxplot represents median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range. **I**) Mosaic plot shows concordance of de novo clustering of six clusters in RAHBT *vs* clusters determined from centroids identified in TBCRC. **J-K**) *ESR1* (ER) and *ERBB2* (HER2) mRNA abundance are highly correlated with protein levels of ER (**J**) and HER2 (**K**), respectively, as measured by MIBI. **L**) *PGR* mRNA abundance was highest in ER_{high} cluster. **M-N**) mRNA abundance of five metabolic genes across the three DCIS subtypes in TBCRC (**M**) and RAHBT (**N**). In both cohorts, the quiescent cluster showed

the lowest mRNA abundance, quantified by Mann-Whitney test. **O**) The quiescent cluster showed the lowest proliferation index in both TBCRC and RAHBT, quantified by Mann-Whitney test. **P**) Kaplan-Meier plot of time to progression stratified by DCIS subtypes in RAHBT.

Figure S4: Characterizing the CNA landscape of DCIS, supplemental to Figure 4.

A) PGA is not correlated with the number of mapped reads. **B**) The 17 recurrent CNAs were not robustly associated with progression. Hazard ratios from CoxPH modeling correcting for treatment. Vertical dotted line represents HR = 1. Covariate on the right indicates if CNA is gain (red) or loss (blue). **C**) PGA (median dichotomized) is not associated with progression. **D**) Consensus matrix from NMF unsupervised clustering of the CNA landscape of DCIS. **E-F**) CNA clusters identified in TBCRC alone (y-axis, **E**) or RAHBT alone (**F**) are highly concordant with those identified considering TBCRC and RAHBT jointly (x-axis). Blue indicates enrichment while red indicates depleted. **G**) CNA clusters are not associated with tumor purity. Boxplot represents median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range. P-value from Kruskal-Wallis test. **H-I**) Kaplan-Meier plot of time to progression stratified by the six CNA clusters in TBCRC (**H**) and RAHBT (**I**). **J**) CNA (y-axis) and RNA (x-axis) clusters are not highly correlated.

Figure S5: Analysis of the tumor microenvironment, supplemental to Figure 5.

A) scRNAseq dataset (Azizi et al. Cell, 2018) used to build the signature matrix. **B**) Signature matrix created with CSx, with 12 different immune cell types. **C**) In-silico validation of signature matrix: A set of samples from the same scRNAseq dataset was reserved to build a synthetic matrix of bulk RNA-seq data. By mixing different proportions of single cell transcripts, the synthetic bulk was used to analyze the correlation between known vs obtained cell proportions by CSx. P-value and R from Pearson's correlation. **D**) Protein validation of CSx signature matrix by MIBI. Correlogram showing MIBI-estimated cell types vs CSx-estimated cell types in RAHBT samples. **E**) Percentage of fibroblasts, endothelial and total immune cells present in each stromal cluster estimated by CSx. **F**) Abundance of total myeloid, lymphoid and granulocyte cells, represented as percentage of total immune cells. **G-H**) Abundance of 12 immune cell types, represented as percentage of total immune cells, by

stromal clusters. Box plots (**E-H**): Red: Normal-like. Blue: Collage rich. Green: Desmoplastic. Purple: Immune dense. *: adj.P <.05; **: adj.P < 0.01; ***: adj.P<0.001; ****:adj.P < 0.0001. **I**) Correlation between 4 stromal clusters and PAM50 classification. **J**) Correlation between 4 stromal clusters and de novo DCIS classification. **K**) Kaplan-Meier analysis of time to iBE by stromal clusters. **L**) Forest plot of the stromal clusters, HR and CI from CoxPH analysis. Cluster 3 had the lowest hazard ratio. **M-O**) Proliferation index (**M**), EMT score (**N**), and PGA (**O**) assessed by stromal clusters. No significant difference was found between clusters. **P**) Heatmap of 300 DE genes in epithelial samples of cluster 4 compared to the rest of the clusters. **Q**) GO analysis of DE genes of epithelial samples classified by stromal subtypes. Immune-related pathways were upregulated in cluster 4 epithelium. **R**) In order to compare DCIS stromal subtypes with normal and IBC stroma, we identified DCIS stromal clusters' centroids and applied them to normal and IBC. Figure S5 R shows a high correlation between predicted and de novo identified normal and IBC stromal clusters. **S**) GO analysis of DE genes after differential abundance analysis (DESeq2, one-vs-rest) comparing four DCIS stromal subtypes with normal stroma (n=10) and IBC stroma (n=30).

Figure S6: Outcome analysis, supplemental to Figure 6.

A) Consensus plot for NMF 812 k=4 solution, TBCRC. **B**) Silhouette plot for NMF 812 k=4 solution, TBCRC. **C**) Mosaic plots, NMF 812 k=4 clusters vs. ER, HER2 status, grade, necrosis, PAM50 and RNA clusters in TBCRC. **D**) Kaplan-Meier analysis, NMF 812 k=4 solution, TBCRC, DCIS iBEs only. **E**) Kaplan-Meier analysis, NMF 812 k=4 solution, TBCRC, IBC iBEs only. **F**) Heatmap of 812 genes and NMF clusters (k=3) in RAHBT. Top bars show ER and HER2 status, PAM50 and novel DCIS cluster classifications, DCIS grade, and necrosis. **G**) Silhouette plot for NMF 812 k=3 solution, RAHBT. **H)** Consensus plot for NMF 812 k=3 solution, RAHBT. **I**) Kaplan-Meier analysis, NMF 812 k=3 solution, RAHBT. **J**) Mosaic plots, NMF 812 k=3 clusters vs. ER, HER2 status, grade, necrosis, PAM50, RNA clusters, and stromal clusters in RAHBT. **K**) Kaplan-Meier analysis, NMF 812 k=3 solution, RAHBT, DCIS iBEs only. **L**) Kaplan-Meier analysis, NMF 812 k=3 solution, RAHBT, IBC iBEs only. **M**) Schematic of model training for RNA, CNA, TME and MV models.

Tables

Table 1. Breast Pre-cancer Atlas Retrospective Patient Cohorts with RNA-seq data

PCA RNA-seq Only	TBCRC				RAHBT						Grand Total
	DCIS without recurrence (N=95)	DCIS with DCIS Recurrence (N=66)	DCIS with Invasive Recurrence (N=55)	TBCRC Total (N=216)	DCIS without recurrence (N=184)	DCIS with Ipsilateral DCIS Recurrence (N=17)	DCIS with Ipsilateral Invasive Recurrence (N=29)	DCIS with Contralateral DCIS (N=19)	DCIS with Contralateral Invasive Disease (N=16)	RAHBT Total (N=265)	(N=481)
Year of Diagnosis											
Median	2009	2008	2006	2008	2002	2005	2000	2002	1991	2002	2006
Age at Diagnosis											
Median	54	54	50	52	53	57	48	57	53.5	53	53
Mean (±SD)	54.4 (±8.5)	55.2 (± 9.8)	52.6 (±9.8)	54.0 (±9.2)	55.6 (±11.4)	58.2 (±12.2)	49.9 (±10.3)	55.9 (±9.9)	58.2 (±12.2)	55.5 (±11.5)	54.8 (±10.6)
Grade											
1	5 [5.3%]	6 [9.0%]	3 [5.5%]	14 [6.5%]	51 [27.7%]	3 [17.6%]	8 [27.6%]	7 [36.8%]	4 [25.0%]	73 [27.5%]	87 [18.15]
2	37 [38.9%]	26 [39.4%]	19 [34.5%]	82 [37.9%]	65 [35.3%]	7 [41.2%]	15 [51.7%]	8 [42.1%]	7 [43.8%]	102 [38.5%]	184 [38.3%]
3	53 [558%]	34 [51.5%]	33 [60.0%]	120 [55.6%]	65 [35.3%]	6 [35.3%]	4 [13.8%]	4 [21.1%]	5 [31.3%]	84 [31.7%]	204 [42.4%]
Missing	0	0	0	0	3 [1.6%]	1 [5.9%]	2 [6.9%]	0	0	6 [2.3%]	6 [1.2%]
Pathologic Tumor Size											
Median	2.1	1.5	1.9	1.9	NA	NA	NA	NA	NA	NA	NA
Mean (±SD)	2.7 (± 1.9)	2.2 (± 2.0)	2.8 (± 2.6)	2.6 (±2.1)	NA	NA	NA	NA	NA	NA	NA
Marker Status											
ER(+)	60 [63.2%]	41 [62.1%]	37 [67.3%]	138 [63.9%]	123 [66.8%]	11 [64.7%]	24 [82.8%]	17 [89.5%]	14 [87.5%]	189 [71.3%]	327 [68.0%]
ER(-)	35 [36.8%]	25 [37.9%]	18 [32.7%]	78 [36.1%]	61 [33.2%]	6 [35.3%]	5 [17.2%]	2 [10.5%]	2 [12.5%]	76 [28.7%]	154 [32.0%]
ER(+) Dx before 2000	0	2 [3.0%]	4 [7.3%]	6 [2.7%]	46 [25.0%]	2 [11.8%]	10 [34.5%]	7 [36.8%]	9 [56.2%]	74 [27.9%]	80 [16.6%]
ER(+) Dx 2000 & after	60 [63.2%]	39 [59.1%]	33 [60.0%]	132 [61.1%]	29 [15.8%]	9 [52.9%]	14 [48.3%]	10 [52.6%]	5 [31.2%]	67 [25.3%]	199 [41.4%]
ER(-) Dx before 2000	0	0	1 [1.8%]	1 [0.5%]	77 [41.8%]	3 [17.6%]	4 [13.8%]	2 [10.5%]	1 [6.3%]	87 [32.8%]	88 [18.3%]
ER(-) Dx 2000 &	35 [36.8%]	25 [37.9%]	17 [30.9%]	77 [35.6%]	32 [17.4%]	3 [17.6%]	1 [3.4%]	0	1 [6.3%]	37 [14.0%]	114 [23.7%]

after											
ER(+) Dx 2000 & after	60 [63.2%]	39 [60.9%]	33 [66.0%]	132 [63.2%]	29 [47.5%]	9 [75.0%]	14 [93.3%]	10 [100.0%]	5 [83.3%]	67 [64.4%]	199 [63.6%]
ER(-) Dx 2000 & after	35 [36.8%]	25 [39.1%]	17 [34.0%]	77 [36.8%]	32 [52.5%]	3 [25.0%]	1 [6.7%]	0	1 [16.7%]	37 [35.6%]	114 [36.4%]
ER(+) Dx before 2000	0	2 [100.0%]	4 [80.0%]	6 [85.7%]	46 [37.4%]	2 [40.0%]	10 [71.4%]	7 [77.8%]	9 [90.0%]	74 [46.0%]	80 [47.6%]
ER(-) Dx before 2000	0	0	1 [20.0%]	1 [14.3%]	77 [62.6%]	3 [60.0%]	4 [28.6%]	2 [22.2%	1 [10.0%]	87 [54.0%]	88 [52.4%]
Treatment											
Lumpectomy w Radiation	58 [61.1%]	40 [60.6%]	22 [40.0%]	120 [55.5%]	91 [49.5%]	12 [70.6%]	18 [62.1%]	8 [42.1%]	8 [50.0%]	17 [51.7%]	257 [53.4%]
Lumpectomy no Radiation	5 [5.3%]	16 [25.2%]	12 [21.8%]	33 [15.3%]	34 [18.5%]	5 [29.4%]	7 [24.1%]	1 [5.3%]	0	47 [17.7%]	80 [16.6%]
Lumpectomy Radiation Unknown	1 [1.1%]	1 [1.5%]	2 [3.6%]	4 [1.9%]	3 [1.6%]	0	1 [3.4%]	1 [5.3%]	1 [6.3%]	6 [2.3%]	10 [2.1%]
Mastectomy	31 [32.6%]	9 [13.6%]	19 [34.5%]	59 [27.3%]	56 [30.4%]	0	3 [10.3%]	9 [47.4%]	7 [43.8%]	75 [28.3%]	134 [27.9%]
Time to Recurrence (months)				N=121						N=81	N=202
Median		40.0	58.0	48.0		49.5	80.1	80.6	55.6	62.3	51.5
Mean (±SD)		52.7 (±39.9)	71.2 (±43.9)	61.1 (±42.6)		61.5 (±43.6)	92.2 (±74.2)	107.3 (±89.1)	71.3 (±56.3)	85.5 (±70.6)	70.9 (±56.7)
Follow Up Time (months)											
Median	92.0	40.0	58.0	74.0	113.1	49.5	80.1	80.6	55.6	100.2	85.2
Mean (±SD)	105.7 (± 37.0)	52.7 (±39.9)	71.2 (±43.9)	80.7 (±45.9)	129.8 (±85.6)	61.5 (±43.6)	92.2 (±74.2)	107.3 (±89.1)	71.3 (±56.3)	116.3 (±83.8)	100.3 (±71.5)
Margins											
Ink on tumor	0	0	0		9 [4.9%]	2 [11.8%]	2 [6.9%]	3 [15.8%]	1 [6.3%]	17 [6.4%]	17 [3.5%]
<2mm	27 [28.4%]	28 [42.4%]	17 [30.9%]	72 [33.3%]	24 [13.0%]	3 [17.6%]	3 [10.3%]	3 [15.8%]	3 [18.8%]	36 [13.6%]	108 [22.5%]
At least 2mm	37 [38.9%]	25 [37.9%]	21 [38.2%]	83 [38.4%]	27 [14.7%]	4 23.5%]	2 [6.9%]	3 [15.8%]	2 [12.5%]	38 [14.3%]	121 [25.2%]
Clear, unknown mm	31 [32.6%]	13 [19.7%]	17 [30.9%]	61 [28.2%]	81 [44.0%]	8 [47.1%]	17 [58.6%]	8 [42.1%]	4 [25.0%]	118 [44.5%]	179 [37.2%]
Missing	0	0	0	0	43 [23.4%]	0	5 [17.2%]	2 [10.5%]	6 [37.5%]	56 [21.1%]	56 [11.6%]
Race											
White	62 [65.2%]	38 [57.6%]	28 [50.9%]	128 [59.3%]	138 [75.0%]	12 [70.6%]	22 [75.9%]	15 [78.9%]	10 [62.5%]	197 [74.3%]	325 [67.6%]
Black	22 [23.2%]	21 [31.8%]	22 [40.0%]	65 [30.0%]	45 [24.5%]	5 [29.4%]	7 [24.1%]	3 [15.8%]	6 [37.5%]	66 [24.9%]	131 [27.2%]
Asian	2 [2.1%]	1 [1.5%]	2 [3.6%]	5 [2.3%]	0	0	0	0	0	0	5 [1.0%]
Pacific Islander	0	1 [1.5%]	0	1 [0.5%]	0	0	0	1 [5.3%]	0	1 [0.4%]	2 [0.4%]
Other	0	0	0	0	0	0	0	0	0	0	0

Unknown	9 [9.5%]	5 [7.6%]	3 [5.5%]	17 [7.9%]	1 [0.5%]	0	0	0	0	1 [0.4%]	18 [3.7%]

STAR Methods

1. Cohort collection and sample acquisition

RAHBT Cohort

The Resource of Archival Breast Tissue (RAHBT) is a data/tissue resource established by Drs. Allred and Colditz in 2008 focused on premalignant or benign breast disease. Uniform coding of premalignant lesions assures greater consistency and use of research. Follow-up through hospital record linkages documents subsequent breast lesions including IBC. The entire study population includes women ages 18 and older with documented cases of premalignant breast disease (including carcinoma in situ).The study was approved by the Washington University in St. Louis Institutional Review Board (IRB ID #: 201707090).

Women were identified as eligible through seven primary sources: Washington University School of Medicine Departmental databases (Surgery, Radiation Oncology, Pathology, and Radiology), and the Siteman Oncology Services Database (local tumor registry), the St. Louis Breast Tissue Repository, and the Women's Health Repository. We reviewed all records, excluded women with cancer prior to qualifying premalignant lesions and identified 1831 unique women with DCIS or DCIS and subsequent recurrence. A common data set with pathologic details, risk factor data, treatment, and unique identifiers was created and used to follow these women for subsequent breast lesions. Centralized pathology review confirmed 174 cases of DCIS with recurrent lesions. For each case (with subsequent ipsilateral or contralateral breast events) we matched two controls who remained free from subsequent breast events based on race, year of diagnosis (+/- 5 years), age at diagnosis (+/- 5 years), and type of definitive surgery (mastectomy or lumpectomy). For each DCIS diagnosis we retrieved slides and blocks for pathology review, secured a whole slide image of each sample, marked for TMA cores, and prepared for laboratory processing. A total of 172 cases and 338 controls were cored for TMAs. Breast pathology review was completed by Drs. Allred, Warrick, DeSchryver, and Veis.

The total number of patients for the full RAHBT cohort was 510 (**Table S1**). The median age at diagnosis was 54, and median year of diagnosis 2001. Time to recurrence with ipsilateral IBC was 84 months, and to diagnosis of ipsilateral DCIS 47 months. For women in the cohort with no iBEs, follow up extended to 132 months, on average. Treatment of initial DCIS ranged from lumpectomy with radiation (approximately half of cases), and no radiation (20%) and mastectomy (30%). The RAHBT cohort was composed of African American women (26%) and white women (74%).

For RAHBT, 265 patients were analyzed by RNA-seq (**Table 1**). The median age at diagnosis was 53, and median year of diagnosis 2002. Time to recurrence with ipsilateral IBC was 80 months, and to diagnosis of ipsilateral DCIS 50 months. For women in the cohort with no iBEs, average follow up extended to 130 months. Treatment of initial DCIS ranged from lumpectomy with radiation (52%), and no radiation (18%) and mastectomy (28%). This subset of the RAHBT cohort was composed of 25% African American women.

TBCRC 038 Cohort

TBCRC 038 is a retrospective multi-center study activated at 12 participating TBCRC (Translational Breast Cancer Consortium) sites, which identified women treated for ductal carcinoma in situ (DCIS) at one of the enrolling institutions between 01/01/1998 and 02/29/2016. The TBCRC and the Department of Defense (DOD) approved this study for the collection of archival tissues. Duke served as the initiating and central site for all data, samples, assays, and analysis. The study was approved by the Duke Health Institutional Review Board (Protocol ID: Pro00068646) as well as the IRB at each participating institution. Individual sites reviewed medical records to identify patients eligible for the study.

Study eligibility criteria included: Women aged 40-75 years at diagnosis of DCIS without invasion; no prior treatment for breast cancer; and definitive surgical excision with no ink on tumor margins and treated with mastectomy, lumpectomy with radiation, or lumpectomy. Cases (patients with subsequent iBEs) were matched 1:1 to controls with at least 5 years of follow up without subsequent iBEs. Matching was based on year of diagnosis (+/-5 years), age at diagnosis (+/- 5 years), and DCIS nuclear grade (high grade vs. non-high grade). All

cases consisted of initial diagnosis of pure DCIS, with ipsilateral recurrence occurring no less than 12 months from date of primary diagnosis. Clinical data, including treatment data, were collected at each site, and standardized data points were entered into a web-based portal. Tumor tissue was collected from FFPE blocks and cut into 5um sections. All slides were scanned and reviewed centrally by a breast pathologist (AH) to confirm the diagnosis. Tumor tissue marked by the pathologist was macrodissected for bulk analysis assays.

The full TBCRC cohort (**Table S1**) includes 221 patients with new DCIS diagnosed between January 1, 1998 and February 29, 2016, ages 40 to 75, with original DCIS block available, who had been treated with mastectomy, lumpectomy with radiation, or lumpectomy alone. DCIS cases included 95 women without iBEs after 5 or more years, 70 with DCIS iBEs, and 56 with IBC iBEs. Median time to IBC iBEs was 59 months and 37 months to DCIS iBE. African American women constituted 30% of this cohort.

The 216 patients from the TBCRC cohort analyzed by RNA-seq (**Table 1**) includes 95 women without iBE after 5 or more years, 66 with DCIS iBEs, and 55 with IBC iBEs. Median time to IBC iBE for this subset was 58 months and 40 months to DCIS iBE. 30% of this subset were African American.

2. Wet lab methods

a. TMA construction

Qualified DCIS or subsequent lesion slides were assembled for pathology review. The research breast pathologist marked the slides for best area to core (1mm) for the carcinoma in situ and later event. The TMAs were designed such that cases/controls were assigned randomly on the map. The Beecher Tissue Arrayer was used to take a core from the patient donor block and place it in the designated area of the recipient TMA block. Slides were then cut for research purposes, and stained H&E and unstained slides were prepared. The TMAs were stored in the St. Louis Breast Tissue Registry Lab at room temperature.

b. Slide cutting

A TMA cutting breakdown was established to include slides for laser capture microdissection (LCM PEN membrane glass slides) sequencing, multiplex protein (MIBI high-purity gold-coated slides) staining and charged glass slides for FISH analysis of the RAHBT TMAs. The order of the slides for the different assays was as follows:

Slide 1-3: FISH/routine IHC – 4 um slices on charged slides

Slide 4-6: RNA/DNA sequencing – 7 um slices on LCM membrane glass slides

Slide 7: MIBI analysis – 4 um slices on gold coated slides

Slide 8-10: FISH/routine IHC – 4 um slices on charged slides

Slide 11-13: RNA/DNA sequencing – 7 um slices on LCM membrane slides

Slide 14: MIBI analysis – 4 um slices on gold coated slides

Slide 15-17: FISH/routine IHC – 4 um slices on charged slides

Slide 18 H&E stained.

c. Digital H&E generation (scanners)

At Washington University School of Medicine, the H&E original slide and TMA slide for RAHBT was imaged (20x) by Aperio AT2 (Leica). ImageScope provides the software for viewing the slides. Images are stored on secure servers in the Dept of Pathology, Washington University School of Medicine.

d. Pathologic analysis and masking

For the TBCRC cohort, whole slide images of the H&E slide made from the block sourced for DNA and RNA was reviewed and scored for grade, presence of necrosis and architecture by a breast pathologist (AH). For the RAHBT cohort, H&E images from the TMAs were used to score for grade, presence of necrosis and architecture by four breast pathologists (DJV, AH, SS, RBW). Areas of DCIS and normal tissue from the RAHBT TMAs were annotated and masked for LCM by two breast pathologists (SS and RBW).

e. LCM

Consecutive sections of tissue microarray blocks were cut and mounted on PEN membrane slides. Slides were dissected immediately after staining on an Arcturus XT LCM System based on the masked areas. Epithelial and stromal sections were dissected separately (**Figure S1**). Each sample adhere to a CapSure HS LCM Cap (Thermo Fisher #LCM0215). After LCM, the cap was sealed in an 0.5 mL tube (Thermo Fisher #N8010611) and stored at −80°C until library preparation. The matching epithelial regions in consecutive slides were dissected for corresponding DNA libraries.

f. smart-3seq

Sequencing libraries were prepared according to the Smart-3SEQ method (Foley et al., 2019) starting from dissected FFPE tissue on an Arcturus LCM HS Cap, except for the unique P5 index and universal P7 primers. Three control samples were added to each library preparation batch and sequence batch to allow batch effect analysis. Libraries were pooled together according to qPCR measurements and prepared according to the manufacturer's instructions with a 1% spike-in of the PhiX control library (Illumina #FC-110-3002) and sequenced on an Illumina NextSeq 500 instrument with a High Output v2.5 reagent kit (Illumina # 20024906),

g. DNA-seq

Genomic DNA was isolated from LCM FFPE cells using PicoPure DNA Extraction kit (Thermo Fisher Scientific # KIT0103). 50ul lysis buffer with Proteinase K were added to each sample and incubated at 65°C overnight. After inactivating proteinase K, the genomic DNA was cleaned up with AMPure XP beads at 3:1 ratio (Beckman Coulter# A63880) and eluted in the 10mM Tris-HCl (pH8.0).

DNA Libraries were constructed with KAPA HyperPlus Kit (Kapa Biosystems #07962428001). Barcode adapters were used for multiplexed sequencing of libraries with SeqCap Adapter Kit A (Kapa Biosystems #7141530001). DNA libraries were amplified by 19 PCR cycles. AMPure XP beads were used for the size selection and cleaning up. DNA libraries were eluted in the 30 µL 10mM Tris-HCl (pH8.0).

Library size distribution was assessed on an Agilent 2100 Bioanalyzer using the DNA 1000 assay and the concentration was measured by Qubit® dsDNA HS Assay Kit (Thermo Fisher Scientific # Q32851). For each lane, 12 samples were pooled and sequenced by Novogene (Sacramento, CA, US) on the Illumina HiSeq Platform, collecting 110G per 275M reads output of paired-end reads of 150 bp length.

h. MIBI

For full details of the MIBI methods, see the companion paper by Risom et al. Briefly, antibodies were conjugated to isotopic metal reporters. Tissues were sectioned (5µm section thickness) from tissue blocks on gold and tantalum-sputtered microscope slides. Imaging was performed using a MIBI-TOF instrument with a Hyperion ion source.

3. Data processing

a. RNA-seq processing

RNA sequencing data was processed with 3SEQtools (https://github.com/jwfoley/3SEQtools). Single-end Illumina FASTQ files were generated from NextSeq BCL files with bcl2fastq (v2.20.0.422) and then aligned to reference hg38 with STAR aligner (v2.7.3a). Samples that did not meet a minimum threshold of uniquely aligned reads were filtered out. The samples in this study averaged 1.11 million uniquely aligned reads. Gene expression matrices of raw and normalized read counts were produced from BAM files with featureCounts (v1.6.4) of the Subread package (v2.4.2) and GENCODE Release 33.

b. DNA-seq processing

Low-pass WGS data were preprocessed using the Nextflow-base pipeline Sarek (Garcia et al., 2020) v2.6.1 with BWA v0.7.17 for sequence alignment to the reference genome GRCh38/hg38 and GATK (McKenna et al., 2010) v4.1.7.0 to mark duplicates and calibration. The recalibrated reads were further processed and filtered for mappability, GC content using the R/Bioconductor quantitative DNA-sequencing (QDNAseq) v1.22.0 with R v3.6.0. For QDNAseq, 50-kb bins were generated from (http://doi.org/10.5281/zenodo.4274556). We kept only autosomal sequences after filtering due to low-depth mappability and GC correction. We used the QDNAseq corrected output and segmented for CN analysis using the circular binary segmentation (CBS) algorithm from DNAcopy R/Bioconductor package v1.60.0. Copy number aberrations were called using CGHcall v2.48.0 (van de Wiel et al., 2007). The R/Bioconductor package ACE v1.4.0 (Poell et al., 2019) was used to estimate purity and ploidy. Proportion of the genome copy number altered (PGA) was calculated based on CNAs with $|\log_2 \text{ratio}| > 0.3$ based on the following:

$$PGA = \frac{number\ of\ bases\ in\ CNA}{total\ number\ of\ bases\ profiled}$$

c. MIBI

Multiplexed image sets were extracted, slide background-subtracted, denoised, and aggregate filtered. Nuclear segmentation was performed using an adapted version of the DeepCell CNN architecture. Single cell data was extracted for all cell objects and area normalized. The FlowSOM R package v1.22.0 (Van Gassen et al., 2015) was used to assign each cell to one of five major cell lineages (tumor, myoepithelial, fibroblast, endothelial, immune). Immune cells were subclustered to delineate B cells, CD4+ T cells, CD8+ T cells, monocytes, MonoDC cells, DC cells, macrophages, neutrophils, mast cells, double-negative CD4−CD8−T cells, and HLADR+ APC cells. Tumor and fibroblast cells were similarly sub clustered to reveal phenotypic subsets. A total of 16 cell populations were quantified and analyzed. For full details of the MIBI methods, see the companion paper by Risom et al.

4. Analyses

a. ER, HER2 status

We called ER and HER2 positivity based on mRNA abundance levels of *ESR1* and *ERBB2*, respectively. We applied a Guassian mixture model with two components using the mclust R package (v5.4.7).

b. PAM50

PAM50 subtypes were called using the genefu v2.22.1 (Gendoo et al., 2016) R package. We compared the PAM50 subtypes called by genefu against subtypes called adjusting for the expected proportion of ER+ samples, as implemented in (Bergholtz et al., 2020). We found both methods to be highly concordant (>96% concordance). We compared the correlation of DCIS and IBC samples to the PAM50 centroids within the genefu R package using Spearman's correlation. We also compared the silhouette widths based on Euclidean distances of the PAM50 subtypes to the de novo DCIS subtypes using the cluster R package (v2.1.1).

c. Differential abundance analyses

Differential gene expression analysis was performed using the R package DESeq2 v1.30.1 (Love et al., 2014) with default options. P-values were adjusted for multiple testing using the Benjamini-Hochberg method. FDR<0.05 was considered significant for all DESeq2 analyses. For comparison of normal vs. DCIS and DCIS vs. IBC, no patient-matched samples were included in the analyses. Reads matrices were VST normalized for downstream analyses.

d. Unsupervised clustering: Non-negative matrix factorization

We identified RNA and CNA based clusters by non-negative matrix factorization using the NMF R package v0.23.0 (Brunet et al., 2004). Each NMF rank was run 30 times to evaluate cluster stability. We comprehensively evaluated 2-10 clusters for each data type and evaluated cluster fit by cophenetic and silhouette values. RNA clusters were first discovered in TBCRC and replicated in RAHBT. We evaluated replication by quantifying the concordance of de novo clusters identified in RAHBT *vs* clusters determined from centroids identified in TBCRC. CNA clusters were discovered in TBCRC and RAHBT jointly and compared against clusters identified in TBCRC and RAHBT individually to ensure robustness.

For clustering using the 812 gene set, we also used the NMF R package v0.23.0. Each NMF rank was run 30 times, and we comprehensively evaluated 2-10 clusters for each data type and evaluated cluster fit by cophenetic and silhouette values. NMF clustering was run individually for each cohort based on expression of the 812 genes.

e. Identification of recurrent CNAs (GISTIC)

Recurrent CNAs were identified from purity-adjusted segment CNA calls from QDNASeq for 406 DCIS samples using GISTIC2 v2.0.23 (Mermel et al., 2011). GISTIC2 was run with the following parameters: -ta 0.3 -td 0.3 -qvt 0.05 -brlen 0.98 -conf 0.95 -armpeel 1 -res 0.01 -rx 0. To ensure CNAs were not biased by sequencing depth, recurrent CNAs significantly associated (FDR < 0.05) with the number of uniquely mapped reads were filtered out.

Associations were quantified by Mann-Whitney test. The number of uniquely mapped reads was determined from samtools flagstat (v1.9).

f. CIBERSORTx

Using single-cell RNA-seq datasets, a breast specific signature matrix was built to resolve proportions of tumor, fibroblasts, endothelial and immune cells from bulk RNA-seq data (**Figure S5A**, (Azizi et al., 2018). scRNAseq data was downloaded from Gene Expression Omnibus database (GEO data repository accession numbers GSE114727, GSE114725). Normalized counts were obtained by using Seurat R package (v3.2.0). The resultant signature matrix contained 3484 genes and allowed to resolve different immune cell types, including B, CD8 T, CD4 T, NKT, NK, mast cells, neutrophils, monocytes, macrophages and dendritic cells (**Figure S5B**). The signature matrix was first *in-silico* validated. In order to test the accuracy of the signature matrix, a set of samples (1/10 of each type) from the same scRNAseq dataset was reserved to build a synthetic matrix of bulk RNA-seq data. By mixing different proportions of single cell transcripts, the synthetic bulk was used to predict cell type proportions and subsequently correlated with the true proportions used to build the synthetic mix. Pearson's coefficient was >0.75 in all the cases, and most >0.9 (**Figure S5C**). The aforementioned matrix was used to deconvolve the LCM RNA-seq samples and to compare CSx-estimated cell abundance with MIBI-identified cell types. Cell abundance between groups was compared by Wilcoxon test followed by Benjamini-Hochberg correction for multiple testing.

g. Shared Nearest Neighbor clustering

LCM stromal samples from RAHBT were classified using the Shared Nearest Neighbor (SNN) clustering method implemented in the Seurat R package (v3.2.0). Data was normalized by negative binomial regression (sctransform R package, v0.3.2, variable.feature.n = "all.genes"). The first 15 principal components were used to identify the

clusters and 16 different resolutions were compared, selecting resolution 0.75 and four clusters as the final solution. Positive markers were selected at a minimum fraction of 0.25 and the resultant gene list was used to further characterize each cluster by gene ontology and KEGG pathway analysis, implemented in clusterProfiler R package (version 3.18.1).

h. Statistical analyses

We used Mann-Whitney U test to compare continuous values between two groups, as specified in the text. We used the Kruskal-Wallis test to compare continuous values between three groups. All statistical analyses were implemented in the R statistical language (v3.6.1). P-values were corrected for multiple hypothesis testing *via* Bonferroni (when <10 independent tests) or Benjamini & Hochberg (when >10 independent tests).

i. Pathway & Gene Set Enrichment Analyses

Gene set enrichment analyses were performed using fgsea R package (v1.12.0) based on the MSigDB Hallmark and Gene Ontology pathways v7.4, (Subramanian et al., 2005). Genes were ranked by their signed adjusted P-values from differential abundance analysis. Pathways were considered enriched if adjusted P-values < 0.05. We evaluated pathway concordance across the DCIS subtypes using a hypergeometric test.

j. Data visualization

Boxplots, heatmaps, scatterplots and barplots were generated using the BoutrosLab.plotting.general R package v6.0.3 (P'ng et al., 2019), or the R packages ggplot2 (v3.3.3, boxplots), corrplot (v0.84, scatterplots), and ComplexHeatmap (v.2.6.2, heatmaps). Volcano plot was generated using EnhancedVolcano (v1.8.0) in R. UMAPs were generated

using the umap (v0.2.7.0) R package with the number of genes indicated in the text. Mosaic plots were generated using the vcd (v1.4.8) R package.

k. Outcome analysis

Associations with time to event were quantified using Cox Proportional Hazard model correcting for treatment as indicated in the text. To standardize follow-up across TBCRC and RAHBT, we censored the follow-up time at 250 months, the maximum follow-up time in TBCRC. Cases with <1 year follow-up were excluded from outcome analyses. Kaplan-Meier plots as implemented in the R packages survival (v3.2.10) and survminer (v0.4.9) were used to visualize outcome differences.

5. Data and Code Availability

All custom code used to analyze data will be made available through a Github repository.
The datasets generated during this study will be made available on the Human Tumor Atlas Network public repository.
De-identified images, including whole slide images of the H&E slide made from the block sourced for DNA and RNA, will be available on the Human Tumor Atlas Network public repository.

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Robert B West (rbwest@stanford.edu).

REFERENCES

ABBA, M. C., GONG, T., LU, Y., LEE, J., ZHONG, Y., LACUNZA, E., BUTTI, M., TAKATA, Y., GADDIS, S., SHEN, J., ESTECIO, M. R., SAHIN, A. A. & ALDAZ, C. M. 2015. A Molecular Portrait of High-Grade Ductal Carcinoma In Situ. *Cancer Res,* 75**,** 3980-90.

ALLINEN, M., BEROUKHIM, R., CAI, L., BRENNAN, C., LAHTI-DOMENICI, J., HUANG, H., PORTER, D., HU, M., CHIN, L., RICHARDSON, A., SCHNITT, S., SELLERS, W. R. & POLYAK, K. 2004. Molecular characterization of the tumor microenvironment in breast cancer. *Cancer Cell,* 6**,** 17-32.

ALLRED, D. C. 2010. Ductal carcinoma in situ: terminology, classification, and natural history. *J Natl Cancer Inst Monogr,* 2010**,** 134-8.

ALLRED, D. C., CLARK, G. M., TANDON, A. K., MOLINA, R., TORMEY, D. C., OSBORNE, C. K., GILCHRIST, K. W., MANSOUR, E. G., ABELOFF, M., EUDEY, L. & ET AL. 1992. HER-2/neu in node-negative breast cancer: prognostic significance of overexpression influenced by the presence of in situ carcinoma. *J Clin Oncol,* 10**,** 599-605.

AMERICAN CANCER SOCIETY. 2019. Breast Cancer Facts & Figures 2019-2020. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf.

AZIZI, E., CARR, A. J., PLITAS, G., CORNISH, A. E., KONOPACKI, C., PRABHAKARAN, S., NAINYS, J., WU, K., KISELIOVAS, V., SETTY, M., CHOI, K., FROMME, R. M., DAO, P., MCKENNEY, P. T., WASTI, R. C., KADAVERU, K., MAZUTIS, L., RUDENSKY, A. Y. & PE'ER, D. 2018. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell,* 174**,** 1293-1308.e36.

BERGHOLTZ, H., LIEN, T. G., SWANSON, D. M., FRIGESSI, A., DAIDONE, M. G., TOST, J., WÄRNBERG, F. & SØRLIE, T. 2020. Contrasting DCIS and invasive breast cancer by subtype suggests basal-like DCIS as distinct lesions. *NPJ Breast Cancer,* 6**,** 26.

BRUNET, J. P., TAMAYO, P., GOLUB, T. R. & MESIROV, J. P. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A,* 101**,** 4164-9.

CALON, A., TAURIELLO, D. V. & BATLLE, E. 2014. TGF-beta in CAF-mediated tumor growth and metastasis. *Semin Cancer Biol,* 25**,** 15-22.

CAMPBELL, M. J., BAEHNER, F., O'MEARA, T., OJUKWU, E., HAN, B., MUKHTAR, R., TANDON, V., ENDICOTT, M., ZHU, Z., WONG, J., KRINGS, G., AU, A., GRAY, J. W. & ESSERMAN, L. 2017. Characterizing the immune microenvironment in high-risk ductal carcinoma in situ of the breast. *Breast Cancer Res Treat,* 161**,** 17-28.

CASASENT, A. K., SCHALCK, A., GAO, R., SEI, E., LONG, A., PANGBURN, W., CASASENT, T., MERIC-BERNSTAM, F., EDGERTON, M. E. & NAVIN, N. E. 2018. Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell,* 172**,** 205-217 e12.

CURTIS, C., SHAH, S. P., CHIN, S. F., TURASHVILI, G., RUEDA, O. M., DUNNING, M. J., SPEED, D., LYNCH, A. G., SAMARAJIWA, S., YUAN, Y., GRÄF, S., HA, G., HAFFARI, G., BASHASHATI, A., RUSSELL, R., MCKINNEY, S., LANGERØD, A., GREEN, A., PROVENZANO, E., WISHART, G., PINDER, S., WATSON, P., MARKOWETZ, F., MURPHY, L., ELLIS, I., PURUSHOTHAM, A., BØRRESEN-DALE, A. L., BRENTON, J. D., TAVARÉ, S., CALDAS, C. & APARICIO, S. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature,* 486**,** 346-52.

DI CESARE, P., PAVESI, L., VILLANI, L., BATTAGLIA, A., DA PRADA, G. A., RICCARDI, A. & FRASCAROLI, M. 2017. The Relationships between HER2 Overexpression and DCIS Characteristics. *Breast J,* 23**,** 307-314.

FOLEY, J. W., ZHU, C., JOLIVET, P., ZHU, S. X., LU, P., MEANEY, M. J. & WEST, R. B. 2019. Gene expression profiling of single cells from archival tissue with laser-capture microdissection and Smart-3SEQ. *Genome Res,* 29**,** 1816-1825.

GARCIA, M., JUHOS, S., LARSSON, M., OLASON, P. I., MARTIN, M., EISFELDT, J., DILORENZO, S., SANDGREN, J., DÍAZ DE STÅHL, T., EWELS, P., WIRTA, V., NISTÉR, M., KÄLLER, M. & NYSTEDT, B. 2020. Sarek: A

portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Res,* 9**,** 63.

GENDOO, D. M., RATANASIRIGULCHAI, N., SCHRÖDER, M. S., PARÉ, L., PARKER, J. S., PRAT, A. & HAIBE-KAINS, B. 2016. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics,* 32**,** 1097-9.

GIL DEL ALCAZAR, C. R., ALEČKOVIĆ, M. & POLYAK, K. 2020. Immune Escape during Breast Tumor Progression. *Cancer Immunol Res,* 8**,** 422-427.

GIL DEL ALCAZAR, C. R., HUH, S. J., EKRAM, M. B., TRINH, A., LIU, L. L., BECA, F., ZI, X., KWAK, M., BERGHOLTZ, H., SU, Y., DING, L., RUSSNES, H. G., RICHARDSON, A. L., BABSKI, K., MIN HUI KIM, E., MCDONNELL, C. H., 3RD, WAGNER, J., ROWBERRY, R., FREEMAN, G. J., DILLON, D., SORLIE, T., COUSSENS, L. M., GARBER, J. E., FAN, R., BOBOLIS, K., ALLRED, D. C., JEONG, J., PARK, S. Y., MICHOR, F. & POLYAK, K. 2017. Immune Escape in Breast Cancer During In Situ to Invasive Carcinoma Transition. *Cancer Discov,* 7**,** 1098-1115.

GORRINGE, K. L., HUNTER, S. M., PANG, J. M., OPESKIN, K., HILL, P., ROWLEY, S. M., CHOONG, D. Y., THOMPSON, E. R., DOBROVIC, A., FOX, S. B., MANN, G. B. & CAMPBELL, I. G. 2015. Copy number analysis of ductal carcinoma in situ with and without recurrence. *Mod Pathol,* 28**,** 1174-84.

GRIMSHAW, M. J., HAGEMANN, T., AYHAN, A., GILLETT, C. E., BINDER, C. & BALKWILL, F. R. 2004. A role for endothelin-2 and its receptors in breast tumor cell invasion. *Cancer Res,* 64**,** 2461-8.

GROEN, E. J., ELSHOF, L. E., VISSER, L. L., RUTGERS, E. J. T., WINTER-WARNARS, H. A. O., LIPS, E. H. & WESSELING, J. 2017. Finding the balance between over- and under-treatment of ductal carcinoma in situ (DCIS). *Breast,* 31**,** 274-283.

GUPTA, P., GUPTA, N., FOFARIA, N. M., RANJAN, A. & SRIVASTAVA, S. K. 2019. HER2-mediated GLI2 stabilization promotes anoikis resistance and metastasis of breast cancer cells. *Cancer Lett,* 442**,** 68-81.

HENDRY, S., PANG, J. B., BYRNE, D. J., LAKHANI, S. R., CUMMINGS, M. C., CAMPBELL, I. G., MANN, G. B., GORRINGE, K. L. & FOX, S. B. 2017. Relationship of the Breast Ductal Carcinoma In Situ Immune Microenvironment with Clinicopathological and Genetic Features. *Clin Cancer Res,* 23**,** 5210-5217.

HESELMEYER-HADDAD, K., BERROA GARCIA, L. Y., BRADLEY, A., ORTIZ-MELENDEZ, C., LEE, W. J., CHRISTENSEN, R., PRINDIVILLE, S. A., CALZONE, K. A., SOBALLE, P. W., HU, Y., CHOWDHURY, S. A., SCHWARTZ, R., SCHÄFFER, A. A. & RIED, T. 2012. 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.

HINSHAW, D. C. & SHEVDE, L. A. 2019. The Tumor Microenvironment Innately Modulates Cancer Progression. *Cancer Res,* 79**,** 4557-4566.

HOUTHUIJZEN, J. M. & JONKERS, J. 2018. Cancer-associated fibroblasts as key regulators of the breast cancer tumor microenvironment. *Cancer Metastasis Rev,* 37**,** 577-597.

HUSSEIN, M. R. & HASSAN, H. I. 2006. Analysis of the mononuclear inflammatory cell infiltrate in the normal breast, benign proliferative breast disease, in situ and infiltrating ductal breast carcinomas: preliminary observations. *J Clin Pathol,* 59**,** 972-7.

HWANG, E. S., DEVRIES, S., CHEW, K. L., MOORE, D. H., 2ND, KERLIKOWSKE, K., THOR, A., LJUNG, B. M. & WALDMAN, F. M. 2004. Patterns of chromosomal alterations in breast ductal carcinoma in situ. *Clin Cancer Res,* 10**,** 5160-7.

JOHNSON, C. E., GORRINGE, K. L., THOMPSON, E. R., OPESKIN, K., BOYLE, S. E., WANG, Y., HILL, P., MANN, G. B. & CAMPBELL, I. G. 2012. Identification of copy number alterations associated with the progression of DCIS to invasive ductal carcinoma. *Breast Cancer Res Treat,* 133**,** 889-98.

KALLURI, R. 2016. The biology and function of fibroblasts in cancer. *Nat Rev Cancer,* 16**,** 582-98.

KILLCOYNE, S., GREGSON, E., WEDGE, D. C., WOODCOCK, D. J., ELDRIDGE, M. D., DE LA RUE, R., MIREMADI, A., ABBAS, S., BLASKO, A., KOSMIDOU, C., JANUSZEWICZ, W., JENKINS, A. V., GERSTUNG, M. & FITZGERALD, R. C. 2020. Genomic copy number predicts esophageal cancer years before transformation. *Nat Med,* 26**,** 1726-1732.

LESURF, R., AURE, M. R., MØRK, H. H., VITELLI, V., LUNDGREN, S., BØRRESEN-DALE, A. L., KRISTENSEN, V., WÄRNBERG, F., HALLETT, M. & SØRLIE, T. 2016. Molecular Features of Subtype-Specific Progression from Ductal Carcinoma In Situ to Invasive Breast Cancer. *Cell Rep,* 16**,** 1166-1179.

LIN, C. Y., VENNAM, S., PURINGTON, N., LIN, E., VARMA, S., HAN, S., DESA, M., SETO, T., WANG, N. J., STEHR, H., TROXELL, M. L., KURIAN, A. W. & WEST, R. B. 2019. Genomic landscape of ductal carcinoma in situ and association with progression. *Breast Cancer Res Treat,* 178**,** 307-316.

LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol,* 15**,** 550.

MA, X. J., SALUNGA, R., TUGGLE, J. T., GAUDET, J., ENRIGHT, E., MCQUARY, P., PAYETTE, T., PISTONE, M., STECKER, K., ZHANG, B. M., ZHOU, Y. X., VARNHOLT, H., SMITH, B., GADD, M., CHATFIELD, E., KESSLER, J., BAER, T. M., ERLANDER, M. G. & SGROI, D. C. 2003. Gene expression profiles of human breast cancer progression. *Proc Natl Acad Sci U S A,* 100**,** 5974-9.

MAK, M. P., TONG, P., DIAO, L., CARDNELL, R. J., GIBBONS, D. L., WILLIAM, W. N., SKOULIDIS, F., PARRA, E. R., RODRIGUEZ-CANALES, J., WISTUBA, II, HEYMACH, J. V., WEINSTEIN, J. N., COOMBES, K. R., WANG, J. & BYERS, L. A. 2016. A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition. *Clin Cancer Res,* 22**,** 609-20.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res,* 20**,** 1297-303.

MEATTINI, I., SAIEVA, C., BASTIANI, P., MARTELLA, F., FRANCOLINI, G., LO RUSSO, M., PAOLETTI, L., DORIA, M., DESIDERI, I., TERZIANI, F., DE LUCA CARDILLO, C., BENDINELLI, B., CIABATTI, C., MUNTONI, C., TINACCI, G., NORI, J., SMITH, H., BRANCATO, B., GALLI, L., SANCHEZ, L. J., CASELLA, D., BERNINI, M., ORZALESI, L., CARTA, G. A., BIANCHI, S., ROSSI, F. & LIVI, L. 2017. Impact of hormonal status on outcome of ductal carcinoma in situ treated with breast-conserving surgery plus radiotherapy: Long-term experience from two large-institutional series. *Breast,* 33**,** 139-144.

MERMEL, C. H., SCHUMACHER, S. E., HILL, B., MEYERSON, M. L., BEROUKHIM, R. & GETZ, G. 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol,* 12**,** R41.

MUSTAFA, R. E., DESTEFANO, L. M., BAHNG, J., YOON-FLANNERY, K., FISHER, C. S., ZHANG, P. J., TCHOU, J., CZERNIECKI, B. J. & DE LA CRUZ, L. M. 2017. Evaluating the Risk of Upstaging HER2-Positive DCIS to Invasive Breast Cancer. *Ann Surg Oncol,* 24**,** 2999-3003.

NEWBURGER, D. E., KASHEF-HAGHIGHI, D., WENG, Z., SALARI, R., SWEENEY, R. T., BRUNNER, A. L., ZHU, S. X., GUO, X., VARMA, S., TROXELL, M. L., WEST, R. B., BATZOGLOU, S. & SIDOW, A. 2013. Genome evolution during progression to breast cancer. *Genome Res,* 23**,** 1097-108.

P'NG, C., GREEN, J., CHONG, L. C., WAGGOTT, D., PROKOPEC, S. D., SHAMSI, M., NGUYEN, F., MAK, D. Y. F., LAM, F., ALBUQUERQUE, M. A., WU, Y., JUNG, E. H., STARMANS, M. H. W., CHAN-SENG-YUE, M. A., YAO, C. Q., LIANG, B., LALONDE, E., HAIDER, S., SIMONE, N. A., SENDOREK, D., CHU, K. C., MOON, N. C., FOX, N. S., GRZADKOWSKI, M. R., HARDING, N. J., FUNG, C., MURDOCH, A. R., HOULAHAN, K. E., WANG, J., GARCIA, D. R., DE BORJA, R., SUN, R. X., LIN, X., CHEN, G. M., LU, A., SHIAH, Y. J., ZIA, A., KEARNS, R. & BOUTROS, P. C. 2019. BPG: Seamless, automated and interactive visualization of scientific data. *BMC Bioinformatics,* 20**,** 42.

PAREJA, F., BROWN, D. N., LEE, J. Y., DA CRUZ PAULA, A., SELENICA, P., BI, R., GEYER, F. C., GAZZO, A., DA SILVA, E. M., VAHDATINIA, M., STYLIANOU, A. A., FERRANDO, L., WEN, H. Y., HICKS, J. B., WEIGELT, B. & REIS-FILHO, J. S. 2020. Whole-Exome Sequencing Analysis of the Progression from Non-Low-Grade Ductal Carcinoma In Situ to Invasive Ductal Carcinoma. *Clin Cancer Res,* 26**,** 3682-3693.

PEROU, C. M., SØRLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H., AKSLEN, L. A., FLUGE, O., PERGAMENSCHIKOV, A., WILLIAMS, C., ZHU, S. X., LØNNING, P.

E., BØRRESEN-DALE, A. L., BROWN, P. O. & BOTSTEIN, D. 2000. Molecular portraits of human breast tumours. *Nature,* 406**,** 747-52.

POELL, J. B., MENDEVILLE, M., SIE, D., BRINK, A., BRAKENHOFF, R. H. & YLSTRA, B. 2019. ACE: absolute copy number estimation from low-coverage whole-genome sequencing data. *Bioinformatics,* 35**,** 2847-2849.

RUEDA, O. M., SAMMUT, S. J., SEOANE, J. A., CHIN, S. F., CASWELL-JIN, J. L., CALLARI, M., BATRA, R., PEREIRA, B., BRUNA, A., ALI, H. R., PROVENZANO, E., LIU, B., PARISIEN, M., GILLETT, C., MCKINNEY, S., GREEN, A. R., MURPHY, L., PURUSHOTHAM, A., ELLIS, I. O., PHAROAH, P. D., RUEDA, C., APARICIO, S., CALDAS, C. & CURTIS, C. 2019. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature,* 567**,** 399-404.

RUSSNES, H. G., VOLLAN, H. K. M., LINGJÆRDE, O. C., KRASNITZ, A., LUNDIN, P., NAUME, B., SØRLIE, T., BORGEN, E., RYE, I. H., LANGERØD, A., CHIN, S. F., TESCHENDORFF, A. E., STEPHENS, P. J., MÅNÉR, S., SCHLICHTING, E., BAUMBUSCH, L. O., KÅRESEN, R., STRATTON, M. P., WIGLER, M., CALDAS, C., ZETTERBERG, A., HICKS, J. & BØRRESEN-DALE, A. L. 2010. Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Sci Transl Med,* 2**,** 38ra47.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. & MESIROV, J. P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A,* 102**,** 15545-50.

SWANSON, D. M., LIEN, T., BERGHOLTZ, H., SØRLIE, T. & FRIGESSI, A. 2019. A Bayesian two-way latent structure model for genomic data integration reveals few pan-genomic cluster subtypes in a breast cancer cohort. *Bioinformatics,* 35**,** 4886-4897.

TRINH, A., GIL DEL ALCAZAR, C. R., SHUKLA, S. A., CHIN, K., CHANG, Y. H., THIBAULT, G., ENG, J., JOVANOVIĆ, B., ALDAZ, C. M., PARK, S. Y., JEONG, J., WU, C., GRAY, J. & POLYAK, K. 2021. Genomic Alterations during the In Situ to Invasive Ductal Breast Carcinoma Transition Shaped by the Immune System. *Mol Cancer Res,* 19**,** 623-635.

TYEKUCHEVA, S., BOWDEN, M., BANGO, C., GIUNCHI, F., HUANG, Y., ZHOU, C., BONDI, A., LIS, R., VAN HEMELRIJCK, M., ANDRÉN, O., ANDERSSON, S. O., WATSON, R. W., PENNINGTON, S., FINN, S. P., MARTIN, N. E., STAMPFER, M. J., PARMIGIANI, G., PENNEY, K. L., FIORENTINO, M., MUCCI, L. A. & LODA, M. 2017. Stromal and epithelial transcriptional map of initiation progression and metastatic potential of human prostate cancer. *Nat Commun,* 8**,** 420.

VAN DE WIEL, M. A., KIM, K. I., VOSSE, S. J., VAN WIERINGEN, W. N., WILTING, S. M. & YLSTRA, B. 2007. CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics,* 23**,** 892-4.

VAN GASSEN, S., CALLEBAUT, B., VAN HELDEN, M. J., LAMBRECHT, B. N., DEMEESTER, P., DHAENE, T. & SAEYS, Y. 2015. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A,* 87**,** 636-45.

VENET, D., DUMONT, J. E. & DETOURS, V. 2011. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol,* 7**,** e1002240.

VINCENT-SALOMON, A., BIDARD, F. C. & PIERGA, J. Y. 2008a. Bone marrow micrometastasis in breast cancer: review of detection methods, prognostic impact and biological issues. *J Clin Pathol,* 61**,** 570-6.

VINCENT-SALOMON, A., LUCCHESI, C., GRUEL, N., RAYNAL, V., PIERRON, G., GOUDEFROYE, R., REYAL, F., RADVANYI, F., SALMON, R., THIERY, J. P., SASTRE-GARAU, X., SIGAL-ZAFRANI, B., FOURQUET, A. & DELATTRE, O. 2008b. Integrated genomic and transcriptomic analysis of ductal carcinoma in situ of the breast. *Clin Cancer Res,* 14**,** 1956-65.

WHELAN, K. A., SCHWAB, L. P., KARAKASHEV, S. V., FRANCHETTI, L., JOHANNES, G. J., SEAGROVES, T. N. & REGINATO, M. J. 2013. The oncogene HER2/neu (ERBB2) requires the hypoxia-inducible factor HIF-1 for mammary tumor growth and anoikis resistance. *J Biol Chem,* 288**,** 15865-77.

YAO, J., WEREMOWICZ, S., FENG, B., GENTLEMAN, R. C., MARKS, J. R., GELMAN, R., BRENNAN, C. & POLYAK, K. 2006. Combined cDNA array comparative genomic hybridization and serial analysis of gene expression analysis of breast tumor progression. *Cancer Res,* 66**,** 4065-78.

Graphical abstract

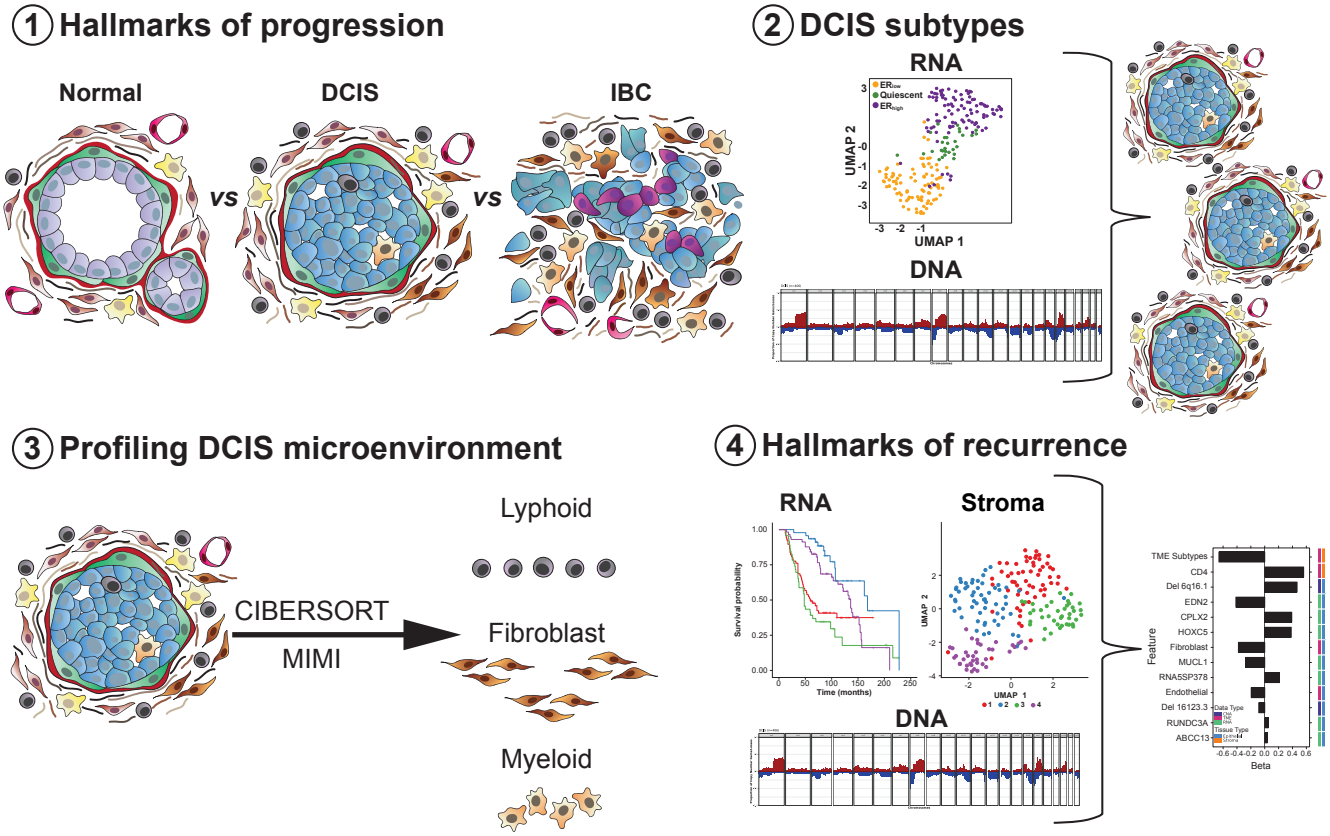

Figure 1

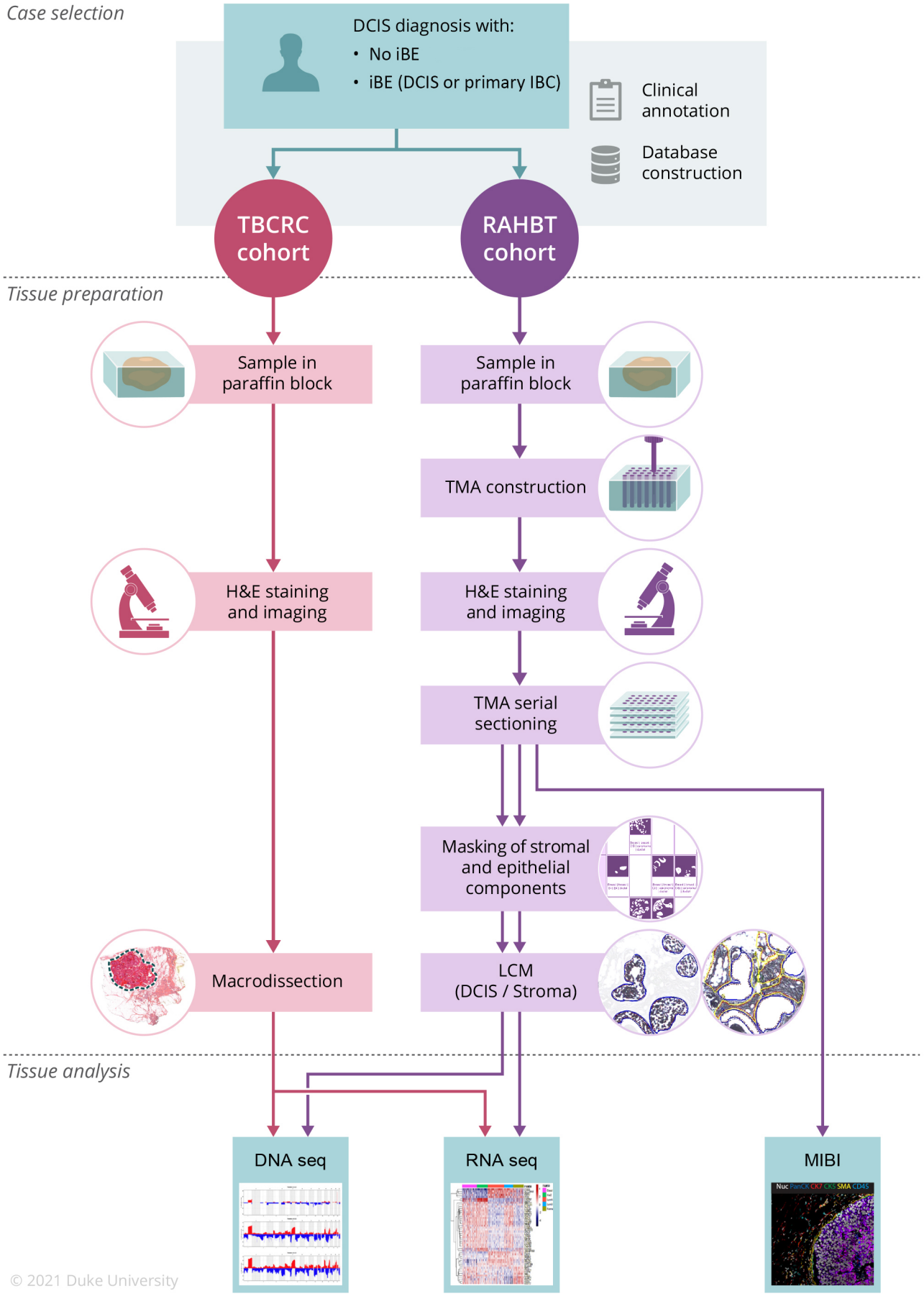

Figure 2

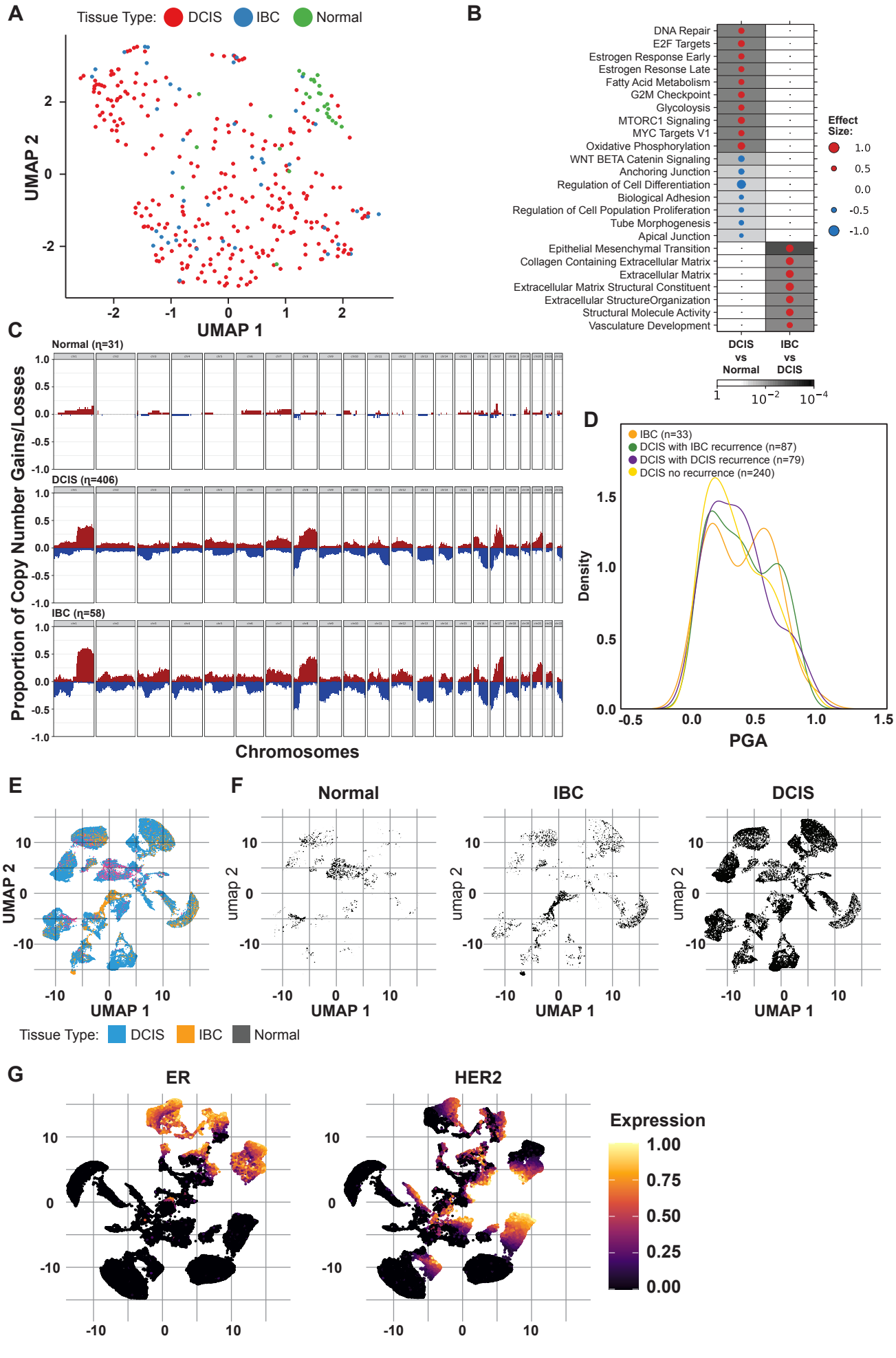

Figure 3

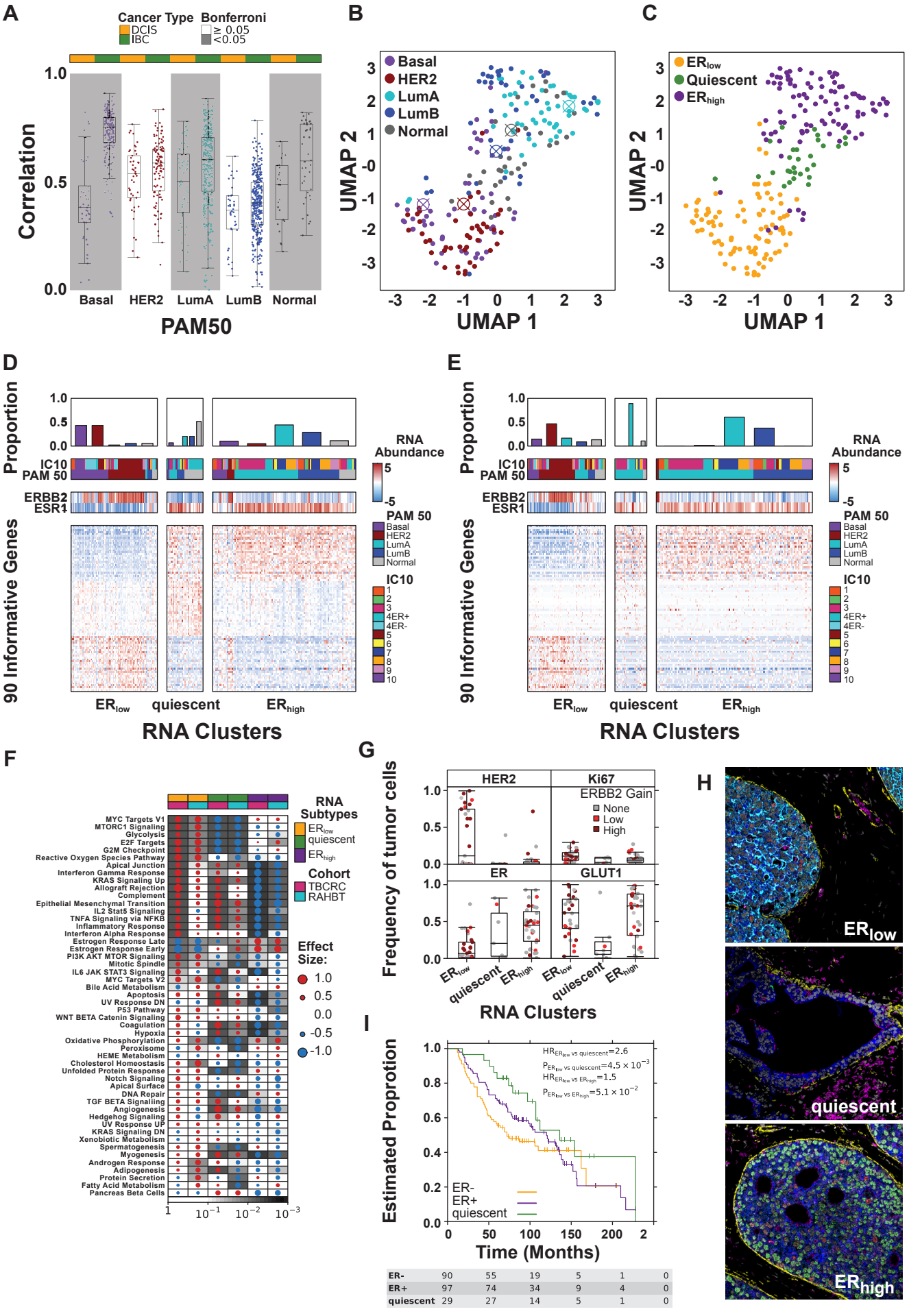

Figure 4

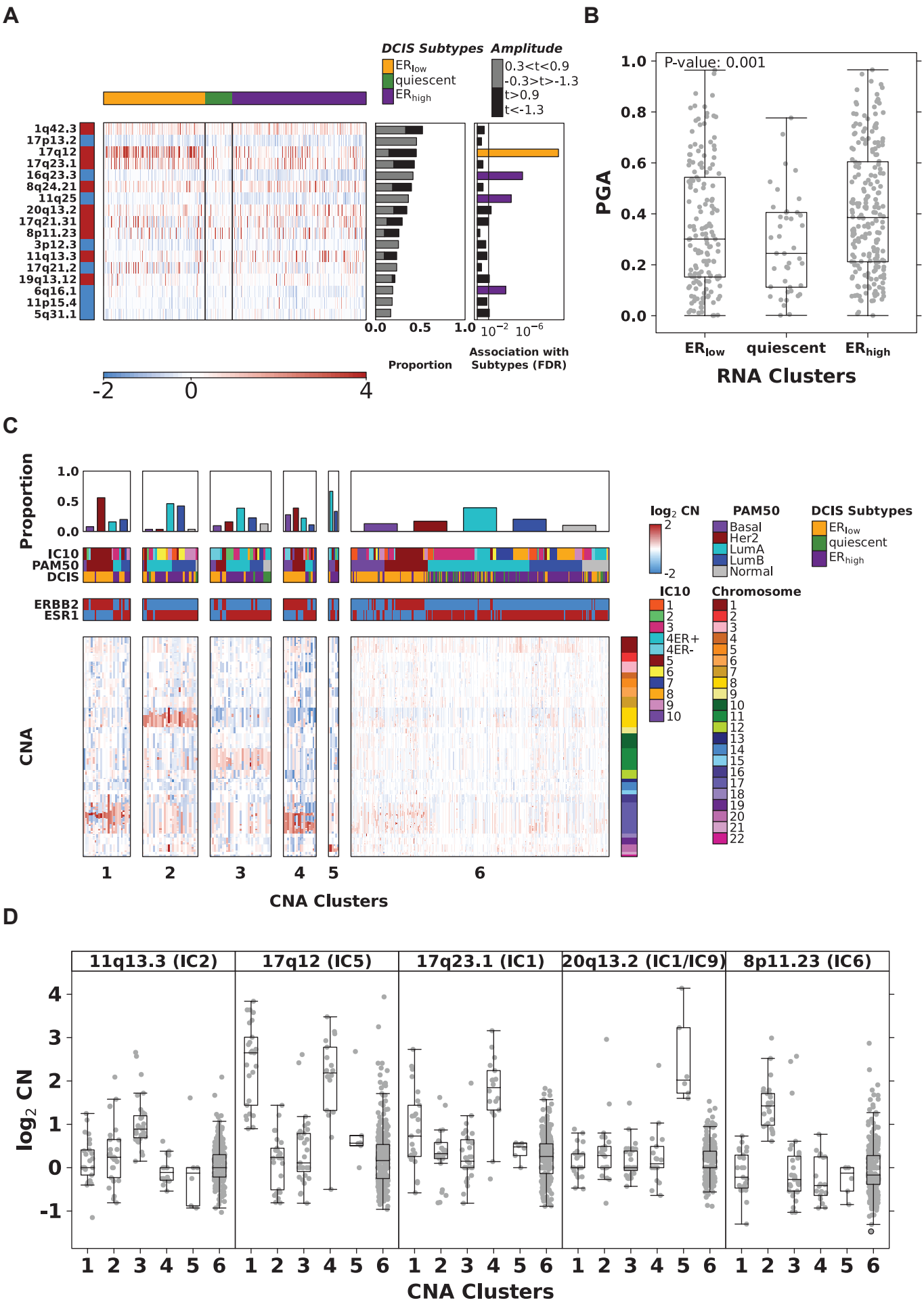

Figure 5

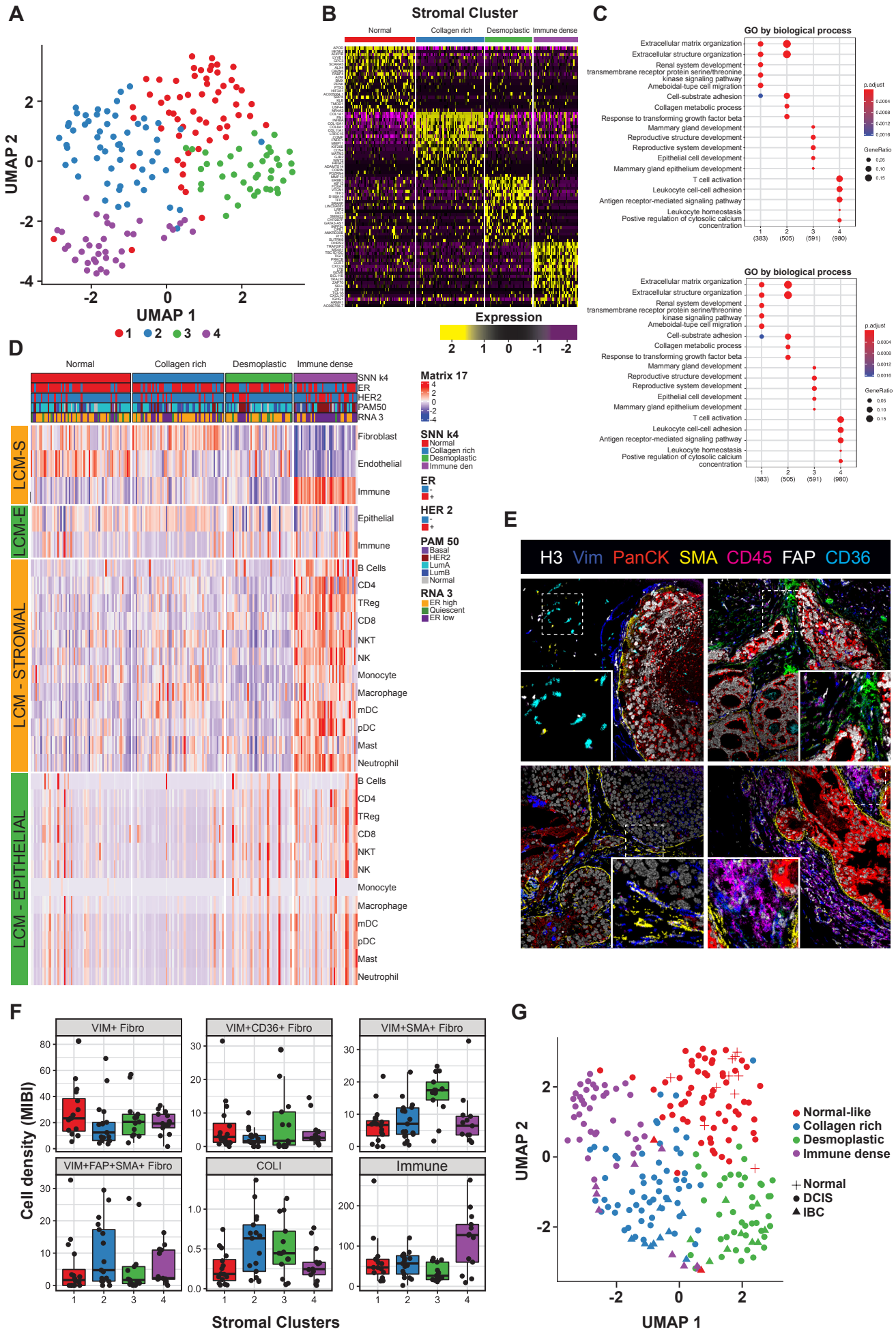

Figure 6