The breast pre-cancer atlas illustrates the molecular and micro-environmental diversity of ductal carcinoma in situ

Daniela Nachmanson[1], Adam Officer[1,2], Hidetoshi Mori[3], Jonathan Gordon[4,5], Mark F. Evans[4,6], Joseph Steward[7] , Huazhen Yao[8], Thomas O'Keefe[9], Farnaz Hasteh[7,10], Gary S. Stein[4,5], Kristen Jepsen[8], Donald L. Weaver[4,6], Gillian L. Hirst[11], Brian L. Sprague[4,12], Laura J. Esserman[11], Alexander D. Borowsky[3], Janet L. Stein[4,5], Olivier Harismendy[2,7, †]

1. Bioinformatics and Systems Biology Graduate Program, University of California San Diego,
2. Division of Biomedical Informatics, Department of Medicine, University of California San Diego,
3. Department of Pathology and Laboratory Medicine, Center for Immunology and Infectious Diseases, School of Medicine, University of California Davis,
4. University of Vermont Cancer Center,
5. Department of Biochemistry, University of Vermont,
6. Department of Pathology and Laboratory Medicine, University of Vermont,
7. Moores Cancer Center, University of California San Diego,
8. Institute for Genomic Medicine, University of California San Diego,
9. Department of Surgery, University of California San Diego,
10. Department of Pathology, University of California San Diego,
11. Helen Diller Family Comprehensive Cancer Center, University of California San Francisco,
12. Department of Surgery, University of Vermont

† Corresponding author: Olivier Harismendy oharismendy@ucsd.edu

Abstract

The increased detection and treatment of early stage breast cancer as well as ductal carcinoma in situ (DCIS) has not led to significant survival benefits. Therefore, the current standard treatment of DCIS is questionable. An informed evidence-based treatment strategy, and likely de-escalation from the current standards requires new prognostic models built from more comprehensive characterization with objective criteria. Parallel profiling of the molecular landscape and micro-environment in pure DCIS remains challenging due to histological heterogeneity and the inevitable reliance on small archived specimens. Leveraging recent methodological advances, we characterized the mutational, transcriptional, histological and microenvironmental landscape across multiple micro-dissected regions from 39 cases to generate a multi-modal breast precancer atlas. The histological architecture was associated with grade, adiposity, and intrinsic expression subtypes. Similar to previous findings, high-grade lesions had higher mutational burden, including *TP53* mutations, while low-grade lesions had more frequent 16q losses and *GATA3* mutations. Multi-region analysis revealed most somatic alterations, including whole genome duplication events, were clonal, but genetic divergence increased with distance between regions. In 7/12 evaluable cases, somatic mutations in putative driver genes affected a subset of regions only. This genetic heterogeneity often accompanied phenotypic heterogeneity and regions with low risk features (Normal-like, Luminal A) occurred earlier than those with high-risk features (Her2-like, Basal or necrosis) according to the phylogenetic analysis. The immune-environment was evaluated using multiplex immuno-histochemistry to measure relative stromal and epithelial densities of B lymphocyte (B-cell), T lymphocyte (T-cell) and regulatory T cells (T-reg) and identify 3 immune-states: Active, Suppressed and Excluded (lower epithelial density). All states included both DCIS and adjacent benign regions, and none associated with intrinsic subtypes. The Excluded state was enriched in high-grade DCIS and, compared to benign areas, more likely acquired in DCIS, showing transcriptional evidence of stronger immune-suppression and possible evasion. The breast pre-cancer atlas therefore reveals correlated levels of phenotypic and genotypic heterogeneity, including at sub-histological resolution. These uniquely integrated observations will help scope future studies, prioritize candidate markers for progression risk modelling and identify functional similarities in precursor lesions from other types of adenocarcinomas.

Introduction

Increasing adoption of breast cancer screening and advances in imaging capabilities have improved our ability to identify breast ductal carcinoma in situ (DCIS). Rarely diagnosed 40 years ago, DCIS now comprises nearly 20% of all breast cancer-related diagnoses [1,2]. Unfortunately, this progress has not resulted in decreased breast cancer mortality. Standard treatment, involving surgical excision often complemented with radiation therapy (in the setting of breast conserving surgery) and endocrine recurrence risk reduction (particularly with ER+ DCIS), therefore constitutes overtreatment, and not without treatment-related consequences for many [2,3]. DCIS progression is particularly difficult to study longitudinally due to the current standard of surgical excision of the lesion and the infrequent progression and/or occurrence of new primary lesions over a long timespan(5-10% after 10 years) [4]. Clinicopathological risk factors such as large size, dense breast, younger age, high pathological grade, presence of comedo-necrosis or Her2 positivity have been associated with increased risk of recurrence, but the resulting predictive models, or those relying on gene expression signatures, are currently insufficient to safely distinguish patients to watch from patients to treat [5].

Contrary to common conceptual models, there is little evidence for the sequential accumulation of somatic alterations during progression from in situ to invasive disease, but rather all invasive breast cancer (IBC) intrinsic subtypes and known driver mutations have been identified in DCIS, albeit at variable prevalence [6–11]. Moreover, both single-cell and bulk studies have shown similar clonal make-up of synchronous invasive and in situ lesions, convoluting the idea that clonal selection drives invasion [11,12]. The role of the immune environment has also been investigated, highlighting the higher lymphocyte infiltration in Her2+ or Triple Negative DCIS, or specific immunological make-up of samples at higher risk of progression [13–15]. Similarly, the role of the basal layer, fibroblasts, adipocytes, other stromal cells or overall extracellular matrix has identified features that are different between DCIS and IBC, likely mediated by chemokine signalling and can be associated with known progression risk factors [16–19]. Their active participation in the malignant transformation of the breast epithelium remains to be established as similar mechanisms are typically involved in normal development, activity and aging of the mammary gland [20].

Progress in our understanding of the processes mediating DCIS onset and progression has been considerably hindered by technical and logistical limitations. Indeed, pure DCIS lesions are commonly small in size, formalin and paraffin embedded (which damages nucleic acids) and can display significant histological heterogeneity [21]. As a consequence, comprehensive molecular and cellular assays and their integrated analysis have seldom been performed in pure DCIS cohorts. Capturing evidence of phenotypic, genetic and cellular heterogeneity, and how they relate to each other is necessary to develop a better spatial, temporal and functional understanding of the mechanisms at play. Recent advances in genome-wide assays, becoming compatible with ever more challenging samples [22–25], have improved our ability to connect histological and molecular observations and enabled such application even to individual microbiopsies from a histological slide of pure DCIS.

Here we describe the combined, parallel histological, molecular and immunological profiling of pre-malignant lesions from 39 patients diagnosed with DCIS, including multiple epithelial micro-biopsies within a subset of samples. The dissection of specific epithelial lesions provided a detailed assessment of the association of their histological architecture with intrinsic subtypes, mutational landscape, driver mutations and immunological states. Multi-region profiling resulted in the inference of clonal relationships, illustrating how genotypes related to phenotypes within a specimen. To our knowledge, our report is the first multi-modal and sub-histological profiling of a cohort of pure DCIS, illustrating spatial heterogeneity and placing diverse states of immune-activity observed in their specific molecular and histological context.

Results

Histological and molecular characterization

We collected a total of 43 specimens (referred to as samples) from 39 patients diagnosed with pure DCIS, including three samples from subsequent DCIS diagnosed between 14 and 70 months after the index DCIS (Figure 1A-B, Table S1). Sixty-nine percent (29/42) of the samples were positive for estrogen receptor (ER) expression and 40% (16/40) had *ERBB2* gene overexpression or amplification (Figure S1A). Each sample was further annotated for grade and histological architecture and the annotations were used to identify regions of interest, guide the micro-biopsies of the epithelial areas and the immuno-histological analysis. On the basis of their studied regions, the cohort consisted of 32 high or intermediate grade DCIS (HG-DCIS), 9 low-grade DCIS (LG-DCIS) and 2 low-grade atypical ductal hyperplasia (ADH). The DCIS regions could be further annotated according to their dominant histological architecture (17 cribriform, 19 solid, 3 mixed, 2 micropapillary) and the presence of necrosis (10 comedo-necrosis, 17 other). LG-DCIS were more frequently of cribriform architecture (8/9), while HG-DCIS were frequently necrotic (25/32). The relative area of adipose tissue in each sample varied between 4 and 91 percent as estimated by segmental classification of the whole slide digital image (Figure 1C, Methods). The lower adipose fraction was associated with higher mammographic breast density (p=0.0067) suggesting the sample histology was representative of the whole breast texture. Interestingly, solid DCIS were associated with a higher adipose fraction (median 69% vs 40%, p=0.008), suggesting a contribution of the breast microenvironment to the growth architecture. Overall, the cohort represents a diverse set of pure in situ lesions identified in absence of any detectable invasive component. The studied samples are enriched for DCIS lesions and specifically annotated for their histological architecture. Each sample was profiled using multiple assays, performed on sequential histological sections (4-7 μm) used for whole transcriptome, whole-exome and spatial immune profiling . Whenever possible, the investigated regions were matched across assays to preserve the spatial information in the analysis and limit the variation due to spatial heterogeneity. Spatial heterogeneity was further addressed in 21 samples for which multiple sub-regions were profiled independently.

The expression of genes was measured using high-throughput sequencing of RNA-seq libraries directly prepared from the micro-biopsied regions [23] (Table S2). The samples were classified according to the PAM50 intrinsic subtypes used for invasive breast cancer (IBC), which identified Basal (N=5), Luminal A (N=10), Luminal B (N=6), Her2-like (N=7) and Normal-like (N=10) samples (Figure 1D). Consistent with IBC classification, Luminal A and B were enriched for samples from ER+ cases, while Her2-like were enriched for Her2+ cases. Similarly, Luminal B and Her2-like were enriched in HG-DCIS, while Luminal A was almost exclusively composed of cribriform LG-DCIS. Luminal A and Normal-like represented closely related classes and together comprised the majority of the samples (20/38), which is not unexpected given the higher fraction of low-grade and pure in-situ lesions in the cohort, in contrast with IBC and previous DCIS expression profiling studies [26,27]. The PAM50 subtype of two independent sub-regions with matching histology and grade was determined in 10 samples, and observed to be discordant in 5 samples (Table S2B), which was associated with larger distances between the regions (Mann Whitney, p=0.005, Figure S1B). Interestingly, matched index and recurrent samples from two patients had at least one region with concordant subtype. Across all samples the distribution of probabilities for each PAM50 subtype likely captures such heterogeneity. Normal-like were truly a mix of Normal and Luminal A, while Her2-like tended to have two main subsets: Her2/Basal and Her2/Luminal B. This suggests that subtypes inferred from bulk analysis, even after epithelial micro-dissection, are frequently the result of a variable mixture of pure subtypes.

Subtype differences in mutational landscape

To determine whether any of the histological or molecular subtypes described above were associated with specific genetic alterations, we characterized their mutational landscape. Whole exome sequencing was carried out on micro-biopsies from 30 samples using a procedure specifically optimized for low amount of damaged DNA [22]. Mutations and copy number alterations (CNA) were identified in 27 and 30 samples, respectively (Table S3). The median copy number burden - or fraction of the genome involved in CNA – was 0.14 and was 2.5 fold higher in HG-DCIS (Mann Whitney, p=0.017, Figure 2A-B). Whole genome doubling (WGD) events were detected in 3/8 eligible samples, all of which were low or intermediate grade cribriform DCIS consistent with its early timing in breast carcinogenesis [28] (Table S3). Consistent with previous studies, loss of 16q (13/30) and 17p (12/30) or gain of 1q (12/30) were among the most frequent chromosomal alterations and, while many events were more frequent in HG-DCIS (Figure 2C, Table S4), these hallmarks were also observed in low-grade or benign lesions, including ADH: (1q gain: 1/9, 16q loss: 7/9, 17p loss: 3/9).

We identified between 74 and 207 coding mutations per sample. The mutational burden was higher in HG-DCIS (Mann Whitney, p=0.003) and Her2-like subtypes (Mann Whitney, p=0.025), recognizing that these categories are overlapping. The HG-DCIS burden (4.4 mut/Mb) was higher than previous reports, possibly due to residual germline variants in our study [11,12]. We identified aging-associated mutational signatures (SBS1 and SBS5) in all samples eligible for analysis (N=13), APOBEC signature (SBS2 and SBS13) on 1 intermediate grade solid DCIS and mismatch repair signature (SBS15 or SBS21) in 3 DCIS of variable grade and architecture (Table S5). The APOBEC signature is therefore more rare in DCIS than IBC (~8% vs >75%), but can be present in premalignant lesions. Interestingly, this sample also displayed clustered mutations (N=3 within 1,416 bp) in chromosome 17q (Figure S2), an APOBEC-driven kataegis site frequently seen in IBC [29]. The most recurrently mutated genes were *PIK3CA* (44%), *TP53* (31%), and *GATA3* (20%), and were all affected by known somatic mutations in breast cancer at similar rates to previous studies of pure DCIS [6,7,9–11] (Figure 2D & Table 1). *TP53* mutations were only found in HG-DCIS and associated with high CNA burden (Mann-Whitney, p=0.018), while *GATA3* mutations were only found in cribriform or ADH histologies and associated with LG-DCIS (Fisher Exact, p=0.005). Interestingly, *GATA3* mutations were identified in larger lesions (Mann Whitney, p=0.038), consistent with a similar observation in invasive cancer and the larger tumor size of *GATA3* mutated xenograft models [30,31]. Another 9 selected genes known to be mutated in IBC were recurrently affected by 18 mutations predicted to be deleterious, 4 of which are known somatic mutations [7]. The result suggests that oncogenic driver mutations are already present at the premalignant stage, including in LG-DCIS (e.g. *SF3B1* c.2098A>G) or ADH (*GATA3* c.925-3_925-2del). This is consistent with previous reports and reports of field effect mutations in normal ducts or benign lesions [32,33], though the contribution of these mutations to the lesion progression remains to be determined.

Genetic heterogeneity and clonal diversity

The histologic assessment and expression profiling have revealed variable levels of phenotypic heterogeneity across the samples. In order to determine whether such heterogeneity is present at the genetic level, we measured genetic heterogeneity in two distinct ways: 1) divergence, which measures the genetic distance between regions of a sample and, 2) clonal relationships, which uses phylogenetic tree construction to establish evolutionary order to genetic alterations (Table S3). We measured divergence by computing a CNA-based score on 19 pairs of histologically matching regions in 11 samples (Table S3, Methods). With no pairs completely independent (max score=0.16), the spatial distance separating the dissected regions was correlated with the extent of their genetic divergence (R^2=0.65, p=0.00017, Figure S3), suggesting that, locally, genetic diversity results from wider range proliferation. Divergence was the lowest in 1 ADH, but was not associated with grade, Her2/ER status, or adipose fraction suggesting that local genetic heterogeneity is not associated with progression risk factors.

More precise clonal relationships between regions were evaluated using phylogenetic analysis in 12 samples, comparing CNA, and mutations when available (Figure 3, Figure S4, Methods). While the majority (88.4%) of CNA were shared across all regions of a sample, 11.6% were private to some regions, as observed in 7/12 samples. Multiple samples (3/12) contained mutations in putative cancer driver genes that were private to one region only. These included known and likely pathogenic mutations in *ATR, PIK3CA, MET, KDM5C,* suggesting that not all driver mutations are acquired early. Interestingly, the three samples with the most private CNA displayed discordant histological architecture or discordant PAM50 subtypes between regions, suggesting that within a sample, genetic and phenotypic differences are linked. Furthermore, in 4/5 samples containing regions with discordant histology and 3/4 with discordant PAM50 subtypes, benign regions or regions with Normal or LumA subtypes appeared earlier than regions with Her2 or Basal subtypes or regions with necrosis. This result suggests that in these samples the lesion progressed with time and acquired histological and molecular features previously associated with increased risk of progression.

Substantial heterogeneity and evolutionary patterns are evident in samples like MCL76_061_16200 (Figure 3A-C), where a region of benign columnar alterations preceded two cribriform regions. While all regions shared a WGD event as well as several arm-level CNA and pathogenic mutations in *GATA3* and *SF3B1,* the cribriform region A acquired private 5q and 8q gains and necrotic features. While this example shows tandem genetic and histological changes as seen across the cohort, it also illustrates that despite occurring earlier, the benign region shares many “driver-like” alterations with both cribriform regions. Furthermore, in another example, despite homogeneous cribriform histologies in regions of MCL76_077_15300 (Figure 3D-F) only one cribriform region lost a copy number of chromosome 8, and presented with Her2 PAM50 subtype as opposed to its Luminal A predecessor. Notably bulk studies have shown chromosome 8 loss to be more frequent in Her2 vs Luminal A breast cancers [34]. Taken all together we illustrate abundant genetic heterogeneity in pure DCIS of all histologies and grades that parallels the levels of phenotypic heterogeneity and often accompanies it, even in regions that are millimeters apart.

Regional differences in the immune micro-environment

To measure the diversity of the immune-landscape and to investigate its potential association with molecular or histological features, we used multiplex immuno-histochemistry (mIHC) to measure the number and density of four cell types - T-cells (CD3+), B-cells (CD20+), T-regs (CD3+/FOXP3+) and epithelial cells (PanCK+) - according to their proliferative status (Ki67+). Both epithelial (PanCK+) and adjacent stromal (PanCK- proximal to epithelium) areas from pre-malignant (N=36 regions across 32 samples) or benign and normal (N=21 across 21 samples) histologies were evaluated. Among pre-malignant regions, the high-grade epithelial areas had lower cell density due to larger cell sizes and frequent central necrosis (median 3.8 vs 6.4 10^3 cells/mm2 p<0.03 – Mann-Whitney). Solid lesions had the highest fraction of proliferating epithelial cells (median 11.5% vs 2.8% p<0.02 - Mann-Whitney, Figure 4A), and interestingly 3/10 HG-DCIS cribriform lesions (2 Her2-like, 1 Luminal B) had markedly higher proliferation. We next classified all regions using non-negative matrix factorization of the stromal and epithelial cell densities, resulting in 3 immune-states characterized by their dominant meta-markers (Figure 4B, Figure S5, Table S7): “Active” (ubiquitous high T-cells), “Suppressed” (ubiquitous low T-cells, high B-cells and T-regs) and “Excluded” (high stromal, low epithelial densities). We verified that regions in Excluded state had higher relative stromal cell density, regions in Active state had denser epithelial T-cells and regions in Suppressed state had denser epithelial T-regs and B-cells (Figure 4C-D). A larger fraction of the benign regions were found in Active (7/21) or Suppressed state (9/21) rather than in Excluded state (4/21) and pre-malignant regions in Excluded state were more likely high grade (7/15 vs 2/17 p=0.049). Interestingly, the immune states of benign and pre-malignant regions were concordant in 12/19 matched cases and discordant in 7 whose lesions were specifically in the Excluded state (Figure 4D). This suggests that the Excluded state may be acquired in response to pre-malignant growth, while other states may be intrinsic to various breast micro-environments. Furthermore, pre-malignant regions in Suppressed state were more likely identified in cases younger than 55 (5/8 vs 4/24 OR=7.6 p=0.02), consistent with the younger age of DCIS patients with infiltrating PD-L1+ lymphocytes [35]. We did not observe any associations between immune states and intrinsic subtype, ER or Her2 status, tumor size, breast density, adipose fraction or DCIS architecture suggesting that they may be independent from traditional histopathological progression risk factors.

In order to identify functional differences between immune-states, we evaluated the differential activity of Hallmark and Reactome processes among the 29 DCIS regions with available gene expression information (Figure S6). Compared to Active and Suppressed states, the Excluded state was associated with upregulation of Type 1 and 2 Interferon response, PD1 signaling and proliferation-related processes as well as the repression of Calcineurin-Calnexin cycle (Figure S6). Noting that the epithelium of DCIS in Excluded state were not completely depleted of infiltrating lymphocytes, the upregulated processes were consistent with the higher expression of *PD1* or *CTLA4* genes in DCIS in Excluded state (Figure 4H), albeit not significant, and suggesting a likely continuum of increasing immuno-suppression from Suppressed to Excluded states. More interestingly, the repression of the Calreticulin-Calnexin cycle was confirmed via single-sample enrichment analysis and showed a progressive repression from Active, to Suppressed, to Excluded states (p=0.022, ANOVA, Figure 4I). This suggests that the export of glyco-proteins - including components of MHC1 complex - via the endoplasmic reticulum, impacts immune-surveillance.

Discussion

There is a compelling requirement for a DCIS atlas that delivers a relatively unbiased, multi-modal perspective of pre-invasive breast cancer. Here, we report the multi-modal profiling of a diverse set of pure DCIS. This comprehensive atlas both confirms previous molecular findings and provides a higher resolution histological and spatial context to interpret them. However, with only 3 known recurrences, the significance of our observations for progression prognosis could not be formally established. Our findings provide a landscape of representative pure DCIS identified in absence of invasive lesions. While some lesions were small, others were quite extended (N=14 larger than 4 cm), which should capture factors that may be associated with robust containment. The cohort therefore spans a variety of clinical, histological, phenotypic and genotypic features. Such variety and contrast is critical to ensure this atlas' utility in designing larger studies, or perhaps providing more cautionary interpretation of observations from cohorts enriched for specific risk factors.

At the heart of our study's innovation was the ability to generate molecular profiles from limited amounts of dissected archival tissue specimen. Similar approaches are used to study clonal expansion in normal tissues [24,25], but generally not performed in parallel for RNA and DNA. Importantly some limitation remains and not all assays were successful. The large variability in success rate was not easy to predict. Likely the age of the specimen, its size, fixation conditions and storage conditions all contribute to success variability which cannot be controlled in a retrospective investigation. Additional limitations are analytical, such as the absence of a matched source of normal DNA from every sample which can result in residual germline variants, perhaps inflating the overall mutation rate observed. The use of adjacent normal or benign tissue can also be problematic and there is ample evidence that they also accumulate somatic mutations [36]. In our study, we clearly identified known breast cancer driver mutations in samples from ADH or other benign alterations. Overall, while some samples are unlikely to ever contain sufficient material for profiling or dissection of adjacent normal, as methodologies evolve and advance, the success rate and data quality will improve to make molecular premalignant profiling more accessible and as routine as is the case in invasive cancer.

Our report contributes to two major advances for understanding pre-malignant lesions. First, we characterized most samples across four important modalities all within a maximum of 50 µm sequentially sectioned tissue. Such advances were enabled by pre-analytical improvements allowing us to reduce the tissue requirement, to include small lesions, and to precisely match regions of interest across each modality: histology, epithelial gene expression, DNA mutations, and immune landscape. As a result we could isolate regions with different histological features that may coexist within a specimen and more confidently establish their association with expression subtypes, clonal heterogeneity or immune state. For example, the integration of histology and expression subtypes showed clear correlation between cribriform architecture and Luminal A subtype. By integrating histology, expression subtype and immune state we showed that some immune-states are found in benign areas and that there is no clear association between immune state and expression subtype. Hence, the depth and interpretability of the analysis is considerably increased by integrating all modalities at the regional level. This has been clearly the case in large cancer studies such as the TCGA, or, more recently through the integrated analysis of histological and somatic features in normal, aging tissues [24,25,37]. While most studies do not typically include immuno-histochemical or other multiplexed spatial analysis, other important advancements in this field in the past year include spatial proteomics used to evaluate the structure of the myoepithelium in DCIS, and spatial transcriptomics used to identify the transcriptional effect of driver mutations in DCIS, representing the emerging frontier of premalignant tissue characterization [10,38]. It is therefore likely that additional spatial profiling compatible with FFPE specimens will bring additional prognostic and mechanistic insights in future DCIS studies.

The other important contribution of our study is the sub-histological analysis to compare regions of interest from the same sample and infer phylogenetic relationships between them. While we determined that the majority of the DCIS samples were classified as Normal-like and Luminal A subtype, typically considered less-aggressive subtypes in breast cancer and reflective of the known precursor stage that DCIS represents, we showed evidence for intrinsic heterogeneity in the PAM50 probabilities, either from the distribution of probabilities within a region or from physically separated regions. This is not entirely surprising as bulk expression subtypes are the result of averaging heterogeneity, similar to glioblastoma subtypes [39] or IBC subtypes [40] from single-cell analysis. Such heterogeneity, especially in DCIS, had been proposed before on the basis of marker staining [41] and our results confirm that it may be rather common. Similarly to the frequency of heterogeneity between region subtypes, we identified evidence of genetic heterogeneity in 7/12 cases, including the presence of private putative driver mutations. This fraction may be an underestimate given the close proximity of many selected pairs. However, the majority of putative genetic drivers, copy number hallmarks and even WGD were clonal, shared by all regions investigated, including a few benign regions. This observation supports evolutionary models derived from invasive cancer, including multi-sample studies, that suggest that most driver mutations occur early followed by a phase of clonal expansion. Similar observations were also made in early multi-regional studies in DCIS [41–43] and studies comparing synchronous DCIS-IBC cases using single-cell sequencing [12], providing further evidence that breast cancer genetic evolution starts in the pre-invasive stage and possibly in benign regions. It is likely that driver alterations may even be present in adjacent histologically normal tissue as observed in field effects studies in normal tissue and benign ducts [36,44]. Such effects support an important contribution of host factors to the initial genetic injury. Hence, unlike previous attempts which were focused on histopathological features, including grade, surgical margins [45,46], future DCIS prognostic models will likely need to be derived from lifetime cancer risk models like GAIL [47] or BOADICEA [48] and incorporate host specific factors, such as polygenic risk scores and reproductive factors, that likely contribute to the DCIS initiation and trajectory.

The immune micro-environment of DCIS has been previously investigated, using both quantification of tumor infiltrating lymphocytes (TIL) and more specific immuno-histochemical approaches and revealed clear quantitative and qualitative variation in lymphocyte infiltration, including higher TIL number and more immunosuppressive features in high risk lesions [49]. Importantly, previous studies in pure DCIS did not quantify stromal and epithelial TILs separately. This distinction may be hard to make in IBC, where both compartments interact at the invasive front and pathologist subjectivity can have a major impact [50,51]. However, this separation can be more clearly established in the analysis of DCIS and was critical in the identification of the Excluded immune state in our atlas. While the Active and Suppressed states have been observed before and could readily be identified in our data, the identification of the Excluded state required the use of an analytical method (NMF) to account for the strong correlations that can exist between TILs type and compartments. The inclusion of adjacent benign areas was also important to interpret the significance of the immune-states, as the Excluded state appeared more likely in reaction to the DCIS growth and increased grade. The

Excluded state exhibited features of immune evasion and could represent a more advanced level of immuno-suppression than the Suppressed state, with the consequence of a topological exclusion from the duct. The downregulation of components of the Calreticulin-Calnexin cycle in the epithelium in Excluded state could impact MHC-I export or maturation providing an evasion mechanism, and contrasting with evasion mediated by MHC-I genetic loss observed in IBC [52,53]. It would be interesting to determine whether the immune states identified can explain the variability of response to local injection of anti-PD1 antibody in DCIS patients, and whether any of the states would elicit, or prevent, the desired ductal infiltration by T-cells [54].

As illustrated by our study and recent advances in the profiling of normal tissues [24,25], histopathology and molecular pathology are becoming more integrated fields, generating deeper and broader datasets at increased cellular and spatial resolution, from the most challenging human samples. Future studies of early transformation and pre-cancer biology such as the one presented here will likely benefit the most from such approaches which capture heterogeneity at scale and can help reconcile analog (optical) and digital (genomics and multiplex) observations. As a result, such multi-dimensional integration may help identify common factors mediating epithelial transformation and progression across multiple glands and organs.

Material and Methods

Sample collection and preparation

FFPE blocks were obtained from UCSD or UVM Pathology Departments after surgical biopsy, excision or mastectomy. The study was reviewed and approved by each institutional review board and they granted a waiver of consent. All specimen blocks were de-identified and sectioned sequentially for the following purpose: Hematoxylin-Eosin (H&E) staining (N=1; 4 µM glass slide), Laser Capture Microdissection (LCM; N=3; 7 µM glass slide coated with polyethylene naphthalate – ThermoFisher #LCM0522), multiplex or regular immunohistochemistry (N≥3 4 µM glass slide) and a final H&E staining (N=1; 4 µM glass slide). The H&E slides were scanned at high resolution and reviewed and annotated by the study pathologist. The LCM slides were stored at -20ºC in an airtight container with desiccant until ready for dissection (1 day to 3 months). The LCM sections were thawed and stained with eosin, sections were kept in xylene and dissected within 2 hours of staining. LCM was performed using the ArcturusXM Laser Capture Microdissection System (ThermoFisher). Matching regions from 6 adjacent sections were collected on capsure Macro Cap (for DNA, N=3 slides) or HS caps (for RNA, N=3 slides), region size, and unambiguous match permitting. Post-dissection, all caps were covered and stored at -20ºC with desiccant. *DNA extraction and QC:* The membrane and adhering tissue were peeled off the caps using a razor blade and the peeled membrane was incubated in proteinase K digestion reaction overnight for 16 h at 56°C to maximize DNA yield after cell lysis The DNA was extracted using the QIAamp DNA Micro Kit (Qiagen) and the elution was done in 20 µL. The extracted DNA was quantified by fluorometry (HS dsDNA kit Qbit – Thermofisher).

RNA-Sequencing and analysis

Library Preparation: RNA sequencing was performed using SMART-3Seq, a 3' tagging strategy specifically designed for degraded RNA directly from FFPE LCM specimen [23]. LCM dissected SMART-3Seq libraries were prepared using the standard protocol for FFPE tissue on Arcturus HS LCM Cap and the individual library SPRI purification option. All FFPE LCM dissected libraries were amplified using 19 PCR cycles during indexing to minimize over-amplification of high abundance mRNAs in each library. Libraries were individually analyzed for size distribution on an Agilent 2200 TapeStation with High Sensitivity D1000 reagent kits to verify average library size of 190 bp and stored at -20 C until sequencing. When all libraries were ready for sequencing, 1 μL of each library was then used to create two library pools used for sequencing and quantified by Qubit 2.0 Fluorometer HS DNA assay. Library pools were sequenced with a 1% PhiX spike-in control library and sequenced on an Illumina HiSeq4000, a run type of single read 75 (SR75) and dual index sequencing.

Transcriptome analysis: Read count data was obtained using a dedicated analysis workflow https://github.com/danielanach/SMART-3SEQ-smk. Briefly, sequencing reads were trimmed using cutadapt 1.18, UMIs were processed using the umi_homopolymer.py script in the SMART-3SEQ tools (https://github.com/jwfoley/3SEQtools), aligned using STAR 2.6.1a, deduplicated using the dedup.py script from https://github.com/jwfoley/umi-dedup and read counts were calculated using featureCounts 1.6.3 [55,56]. Count data was then merged and filtered to remove samples with fewer than 55,000 counts and genes with fewer than 10 read counts across all samples. Filtered count data was then loaded into Seurat version 3.2.3 and processed using the *SCTransform()* function version 0.3.2 to regress out the high mitochondrial content variability across the samples [57]. Batch correction was then performed using ComBat to remove variation attributable to the sequencing center (UCSD vs UVM) [58]. PAM50 subtype probabilities were calculated from the *SCTransform* and batch normalized data using the genefu package [59]. Gene set enrichment analysis (GSEA) was performed as in [60] and single-sample GSEA as in [61]. Gene sets from the REACTOME and Hallmark collections in MSigDB were used to compare the excluded to the non-excluded groups, a permutation test was performed to assess the significance of the GSEA results [62,63]. ANOVA was used to compare the ssGSEA results between the three mIHC groups. FDR of less than 0.1 and p-values of less than 0.05 were considered significant.

Whole exome sequencing and primary analysis

LIbrary preparation: DNA was sheared down to 200 base pairs (bp) using Adaptive Focused Acoustics on the Covaris E220 (Covaris Inc) following manufacturer recommendations with 10 μL Low EDTA TE buffer supplemented with 5 μL of truSHEAR buffer using a microTUBE-15. Libraries were prepared using the Accel-NGS 2S PCR-Free DNA Library Kit (Swift Biosciences).

Ligated and purified libraries were amplified using KAPA HiFi HotStart Real-time PCR 2X Master Mix (KAPA Biosystems). Samples were amplified with 5 µL of KAPA P5 and KAPA P7 primers. The reactions were denatured for 45 seconds (sec) at 98°C and amplified 13-15 cycles for 15 sec at 98°C, for 30 sec at 65°C, and for 30 sec at 72°C, followed by final extension for 1 min at 72°C. Samples were amplified until they reached Fluorescent Standard 3, cycles being dependent on input DNA quantity and quality. PCR reactions were then purified using 1x AMPure XP bead clean-up and eluted into 20 µL of nuclease-free water. The resulting libraries were analyzed using the Agilent 4200 Tapestation (D1000 ScreenTape) and quantified by fluorescence (Qubit dsDNA HS assay).

Capture and Sequencing: Samples were paired and combined (12 µL total) to yield a capture “pond” of at least 350 ng, and supplemented with 5 µL of SureSelect XTHS and XT Low Input Blocker Mix. The hybridization and capture was performed using the Human All Exon V7 panel (S31285117) paired with the Agilent SureSelect XT HS Target Enrichment Kit following manufacturer’s recommendations. Post-capture amplification was performed on the beads in a 25 µL reaction: 12.5 µL of nuclease-free water, 10 µL 5x Herculase II Reaction Buffer, 1 µL Herculase II Fusion DNA Polymerase, 0.5 µL 100 millimolar (mM) dNTP Mix and 1 µL SureSelect Post-Capture Primer Mix. The reaction was denatured for 30 sec at 98°C, then amplified for 12 cycles of 98°C for 30 sec, 60°C for 30 sec and 72°C for 1 min, followed by an extension at 72°C for 5 minutes and a final hold at 4°C. Libraries were purified with a 1x AMPure XP bead clean up and eluted into 20 µL nuclease free water in preparation for sequencing. The resulting libraries were analyzed using the Agilent 4200 Tapestation (D1000 ScreenTape) and quantified by fluorescence (Qbit – ThermoFisher). All libraries were sequenced using the HiSeq 4000 sequencer (Illumina) for 100 cycles in Paired-End mode. Libraries with distinct indexes were pooled in equimolar amounts. The sequencing and capture pools were later deconvoluted using program bcl2fastq [19].

Sequencing reads processing and coverage quality control: Sequencing data was analyzed using bcbio-nextgen (v1.1.6) as a workflow manager [20]. Samples prepared with identical targeted panels were down-sampled to have equal number of reads using seqtk sample (v1.3) [21]. Adapter sequences were trimmed using Atropos (v1.1.22), the trimmed reads were subsequently aligned with bwa-mem (v0.7.17) to reference genome hg19, then PCR duplicates were removed using biobambam2 (v2.0.87) [64–66]. Additional BAM file manipulation and collection of QC metrics was performed with picard (v2.20.4) and samtools (v1.9) [67]. The summary statistics of the sequencing and coverage results are presented in Table S8.

Identification of somatic mutation and copy number alterations

Variant calling: Single nucleotide variants (SNVs) and short insertions and deletions (indels) were called with VarDictJava (v1.6.0), and Mutect2 (v2.2) [68,69]. Variants were required to fall within a 10 bp boundary of targeted regions that overlapped with RefSeq genes (v 109.20190905). A pool of normal DNA was created using whole exome sequencing data of blood of 18 unrelated individuals and was used to eliminate artifacts and common germline variants. Only variants called by both algorithms were considered. These variants were then subjected to an initial filtering step with default bcbio-nextgen tumor-only variant calling filters and the following parameters were used: position covered by at least 5 reads, mapping quality more than 45, mean position in read greater than 15, number of average read mis-matches less than 2.5, microsatellite length less than 5, tumor log odds threshold more than 10, Fisher strand bias Phred-scaled probability less than 10 and VAF more than 0.1 [70]. Functional effects were predicted using SnpEff (v4.3.1) [71]. All samples were re-evaluated for the presence of COSMIC (v91) database mutations which have been previously observed in at least 15 patients and fall within 137 known breast cancer driver genes (Table S9) [7].

Germline variant filtering: In absence of matched normal tissue for DCIS samples, somatic mutations were prioritized computationally using the approach from the bcbio-nextgen tumor-only configuration then additionally subjected to more stringent filtering [70]. Briefly, common variants (MAF>10^{-3} or more than 9 individuals) present in population databases - 1000 genomes (v2.8), ExAC (v0.3), or gnomAD exome (v2.1) - were removed unless in a tier 1 gene from the cancer gene consensus and present in either COSMIC (v91) or clinvar (20190513) [7,72–75]. Variants were removed as likely germline if found at a variant allelic fraction (VAF) greater or equal to 0.9 in non-LOH genomic segments – as determined by CNA analysis (below). Lastly, variants were also removed as potential germline (or artifact) if found in more than two patients in the pool of normal (described above).

Single-sample CNA calling: CNVkit [76] was used for calling somatic copy number alterations (CNA) to measure both overall CNA burden, arm and gene level CNA and identify LOH as previously described in [22]. Allele-specific copy number calling algorithm, ASCAT, was used on a select number of samples for which there was sufficient coverage and the algorithm converged on a solution, in order to identify whole-genome doubling events as well as confirm CNA identified by CNVkit [77]. Default parameters were used with ASCAT with the exception of a segmentation penalty of 100 and a gamma of 1.

Multi-region CNA segmentation: To generate harmonized segmentation breakpoints between regions belonging to the same sample, multi-region segmentation as performed with the R CopyNumber (v1.26.0) package [78]. Outliers in CNVkit bin-level log2 copy ratios were detected and modified using Median Absolute Deviation Winsorization with the winsorize() function, segments were then called using the *multipcf()* function with a gamma of 40.

Mutational signatures: Mutational signatures were called on merged region samples using a single-sample variation of SigProfiler with default parameters to decompose into known single-base substitutions (SBS) reported in COSMIC [79,80].

Analysis of the clonal evolution and genetic heterogeneity

*Measurement of genetic divergence :*Divergence was measured on each pair of related regions, a and b, using the following equation:

$$Divergence_{a,b} = \sum_{k=0}^{n} \left| copy\ ratio\ a_k - copy\ ratio\ b_k \right| * \left(\frac{bins_k}{total\ bins} \right)$$

Where k is the copy number segment, $n$is the total segments, and $bins$is the number of bins covered by a segment from the CNVkit input file. The $\frac{bins_k}{total\ bins}$ term was used as a weighted correction factor for the number of bins contributing to a segment. For samples with more than 2 regions, the maximum divergence between any two regions was used to represent the sample.

CNA-based phylogenetic reconstruction: Construction of phylogenetic trees was performed similarly to the methodology outlined in [81]. Briefly, for each sample the log2 copy ratios from multi-sample copy number segments with at least 12 probes (see above), were translated into a matrix containing -1 for loss (log2 copy ratio<-0.6), 0 for neutral (-0.4<= log2 copy ratio <=0.3) and undetermined for anything else. This matrix was then used to generate Maximum Parismony trees using phangorn using default parameters [82].

Mutation-based phylogenetic reconstruction: To allow the analysis of clonal relationships between regions of the same sample, the coverage depth of each allele at any remaining mutated position in any region was extracted using Mutect2 joint variant caller on the sets of aligned reads from each region. In order to call a mutation either absent or present in a region, we used a Bayesian inference model specifically designed for multi-region variant calling [83]. Treeomics (v1.7.10) was run with the default parameters except for e=0.02. The tree solution which matched the CNA-based reconstruction was then integrated into a single tree for Figures 3 and S4.

Multiplex Immuno-histochemistry

Staining: Tissue sections were prepared from formalin-fixed paraffin embedded tissue blocks and cut to 4 micrometers serial sections and mounted on Superfrost Plus (VWR). The procedure for multiplex immunohistochemistry (mIHC) was followed by a manufacturer's protocol for Opal7-color automation IHC kit (Akoya Bioscience), and the staining was performed with Autostainer DISCOVERY ULTRA (Ventana). Antibodies used in mIHC are anti-CD3 (clone 2GV6, Ventana), anti-CD20 (clone L26, Ventana), anti-Ki67 (clone 30-9, Ventana), anti-FOXP3 (clone SP97, Spring), anti-pan cytokeratin (CK; clone AE1/AE3, DAKO), anti-CD117 (clone c-kit, DAKO). The molecular markers of immune panel (CD3, CD20, Ki67, CKs, FOXP3 and CD117) were visualized with Opal520, Opal540, Opal570, Opal620, Opal650 and Opal690, respectively. DAPI counterstaining was performed with Discovery QD DAPI (Roche). ProLong Diamond Antifade Mounting (ThermoScientific) was used for mounting the coverslip. Detailed staining conditions and autostainer's protocols are reported in our recent report [84].

Visualization and analysis: Tissue samples stained with mIHC were scanned with multispectral imaging microscopy (Vectra 3, Akoya Bioscience). Scanned multispectral images were unmixed on inForm software (ver.2.4.0, Akoya Bioscience) to acquire the fluorescence signal from each marker [84]. Imaging analysis was performed on inForm software by identifying tumor (CK+ area) and stroma (CK- area proximal to the epithelium), each nucleated cell and its cell type.

Alternatively, QuPath software [85] was also used to perform similar imaging analysis on unmixed images converted to multi-layered TIFF format by inForm software [84]. Scanned image areas were aggregated into up to three histological regions per sample: main pre-invasive lesion, alternate pre-invasive lesion, benign or normal epithelium. In each region, the stromal and epithelial densities of each cell type and state was calculated, including when cells were not present (density=0). Regularized marker densities into distribution deciles were then used to classify samples using non-negative matrix factorization (Table S7). The immune-states were assigned and named after the hierarchical clustering of the H matrix (meta-marker values).

Whole slide image digital analysis

High resolution whole slide images of H&E stains were loaded into a QuPath (v2.3) project [85]. One analysis area was defined for each specimen, avoiding location of biopsies as well as dust or marked areas. The analysis areas were segmented into superpixels (sigma=5 μm, spacing=50 μm, maxIterations=10, regularization=0.25) and each superpixel was annotated with both Hematoxylin and Eosin Intensity features (size=2 μm, tile size=25 μm). The mean, median, min, max and standard deviation values were then smoothed (Haralick distance=1, Haralick bins=32). Multiple training areas were annotated from each of the following classes: adipose, stroma, inflammation, epithelium (normal and atypical), void, necrosis, blood vessels. Multiple areas across 2 to 4 samples were used to train a Random Tree classifier. The classifier was then applied to all superpixels included in the analysis area. The accuracy of the classifier was assessed both visually and with multiple test areas for each class. Superpixels of the same class were merged into single annotations and the resulting areas recorded. Separate classifiers were used for images from different institutions, to mitigate possible variation staining, scanning or image format. The fraction of adipose area was compared to breast density using Mann-Whitney test comparing dense & heterogeneously dense breast to other lower densities, or comparing solid DCIS to non-solid DCIS lesions.

Data Availability

The raw RNA and DNA sequencing data has been deposited in dbGAP phs002225. High resolution whole slide images of the H&E stains and corresponding annotations can be viewed on the JPL LabCAS portal (digital object identifiers included in the Table S1).

Acknowledgments

We are grateful to Drs. Michael Campbell, Christina Yau and Kathleen Curtius for helpful conversations, Drs. Alfredo Molinolo, Oluwole Fadare, Sharmeela Kaushal, Valeria Estrada and Mrs Kimberly Mcintyre for their support and assistance in the tissue collection, preparation and dissection, Mrs Eliza Jeong, Marcy Andersen and Nicole Lee for their assistance retrieving clinical information. We thank the technical assistance of the Vermont Integrative Genomics Resource Massively Parallel Sequencing Facility with the combined support of the University of Vermont Cancer Center, Lake Champlain Cancer Research Organization, UVM College of Agriculture and Life Sciences, and the UVM Larner College of Medicine. We acknowledge the work of all working groups from the NCI Consortium for the Molecular Characterization of Screen Detected Lesions (MCL), in particular Christopher Amos, Daniel Crichton, Heather Kincaid, Kirsten Anton, Luca Cinquini, David Liu for their assistance in the data management and sharing. Figure 1 was created with BioRender.com and printed with permission.

Funding

This work is supported by funding from the National Institute of Health (U01CA196406, U01CA196406-03S1, U01CA196383, R01DE026644, T32GM008806, T15LM011271), the National Cancer Institute (P30CA023100) and the California Tobacco Related Disease Research Program pre-doctoral fellowship to DN (28DT-0011). The funding bodies had no role in the design of the study; collection, analysis, and interpretation of data; or in the writing of the manuscript.

Conflict of Interest

The authors declare no conflict of interest

Authors Contributions

JLS, GS, GLH, LJE, SDB, OH designed the study, FH, AM, SDB, DLW, BLS, TOK selected and annotated the specimen. JS, KJ, HZ, HM, MFE, JG generated the data, OH, DN, AO, HM analyzed the data, OH and DN wrote the manuscript. All authors reviewed and approved the manuscript.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020;70:7–30.

2. Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. N Engl J Med. 2012;367:1998–2005.

3. Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. Lancet. 2012;380:1778–86.

4. Sprague BL, Vacek PM, Herschorn SD, James TA, Geller BM, Trentham-Dietz A, et al. Time-varying risks of second events following a DCIS diagnosis in the population-based Vermont DCIS cohort. Breast Cancer Res Treat. 2019;174:227–35.

5. Gorringe KL, Fox SB. Ductal Carcinoma In Situ Biology, Biomarkers, and Diagnosis. Front Oncol. 2017;7:248.

6. Pang J-MB, Savas P, Fellowes AP, Mir Arnau G, Kader T, Vedururu R, et al. Breast ductal carcinoma in situ carry mutational driver events representative of invasive breast cancer. Mod Pathol. 2017;30:952–63.

7. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019;47:D941–7.

8. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. American Society of Clinical Oncology (ASCO); 2009;27:1160–7.

9. Lin C-Y, Vennam S, Purington N, Lin E, Varma S, Han S, et al. Genomic landscape of ductal carcinoma in situ and association with progression. Breast Cancer Res Treat. 2019;178:307–16.

10. Nagasawa S, Kuze Y, Maeda I, Kojima Y, Motoyoshi A, Onishi T, et al. Genomic profiling reveals heterogeneous populations of ductal carcinoma in situ of the breast. Commun Biol. 2021;4:438.

11. Pareja F, Brown DN, Lee JY, Da Cruz Paula A, Selenica P, Bi R, et al. Whole-Exome Sequencing Analysis of the Progression from Non-Low-Grade Ductal Carcinoma In Situ to Invasive Ductal Carcinoma. Clin Cancer Res. 2020;26:3682–93.

12. Casasent AK, Schalck A, Gao R, Sei E, Long A, Pangburn W, et al. Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. Cell. 2018;172:205–17.e12.

13. Gerdes MJ, Gökmen-Polar Y, Sui Y, Pang AS, LaPlante N, Harris AL, et al. Single-cell heterogeneity in ductal carcinoma in situ of breast. Mod Pathol. 2018;31:406–17.

14. Pruneri G, Lazzeroni M, Bagnardi V, Tiburzio GB, Rotmensz N, DeCensi A, et al. The prevalence and clinical relevance of tumor-infiltrating lymphocytes (TILs) in ductal carcinoma in situ of the breast. Ann Oncol. 2017;28:321–8.

15. Campbell MJ, Baehner F, O'Meara T, Ojukwu E, Han B, Mukhtar R, et al. Characterizing the

immune microenvironment in high-risk ductal carcinoma in situ of the breast. Breast Cancer Res Treat. 2017;161:17–28.

16. Allen MD, Marshall JF, Jones JL. αvβ6 Expression in myoepithelial cells: a novel marker for predicting DCIS progression with therapeutic potential. Cancer Res. 2014;74:5942–7.

17. Delort L, Cholet J, Decombat C, Vermerie M, Dumontet C, Castelli FA, et al. The Adipose Microenvironment Dysregulates the Mammary Myoepithelial Cells and Could Participate to the Progression of Breast Cancer. Front Cell Dev Biol. 2020;8:571948.

18. Allinen M, Beroukhim R, Cai L, Brennan C, Lahti-Domenici J, Huang H, et al. Molecular characterization of the tumor microenvironment in breast cancer. Cancer Cell. 2004;6:17–32.

19. Hu M, Yao J, Carroll DK, Weremowicz S, Chen H, Carrasco D, et al. Regulation of in situ to invasive breast carcinoma transition. Cancer Cell. 2008;13:394–406.

20. Unsworth A, Anderson R, Britt K. Stromal fibroblasts and the immune microenvironment: partners in mammary gland biology and pathology? J Mammary Gland Biol Neoplasia. 2014;19:169–82.

21. Sinha VC, Piwnica-Worms H. Intratumoral Heterogeneity in Ductal Carcinoma In Situ: Chaos and Consequence. J Mammary Gland Biol Neoplasia. 2018;23:191–205.

22. Nachmanson D, Steward J, Yao H, Officer A, Jeong E, O'Keefe TJ, et al. Mutational profiling of micro-dissected pre-malignant lesions from archived specimens. BMC Med Genomics. 2020;13:173.

23. Foley JW, Zhu C, Jolivet P, Zhu SX, Lu P, Meaney MJ, et al. Gene expression profiling of single cells from archival tissue with laser-capture microdissection and Smart-3SEQ. Genome Res. 2019;29:1816–25.

24. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science. 2015;348:880–6.

25. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age. Science. 2018;362:911–7.

26. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.

27. Lesurf R, Aure MR, Mørk HH, Vitelli V, Oslo Breast Cancer Research Consortium (OSBREAC), Lundgren S, et al. Molecular Features of Subtype-Specific Progression from Ductal Carcinoma In Situ to Invasive Breast Cancer. Cell Rep. 2016;16:1166–79.

28. Bielski CM, Zehir A, Penson AV, Donoghue MTA, Chatila W, Armenia J, et al. Genome doubling shapes the evolution and prognosis of advanced cancers. Nat Genet. 2018;50:1189–95.

29. D'Antonio M, Tamayo P, Mesirov JP, Frazer KA. Kataegis Expression Signature in Breast Cancer Is Associated with Late Onset, Better Prognosis, and Higher HER2 Levels [Internet].

Cell Reports. 2016. p. 672–83. Available from: http://dx.doi.org/10.1016/j.celrep.2016.06.026

30. Afzaljavan F, Sadr AS, Savas S, Pasdar A. GATA3 somatic mutations are associated with clinicopathological features and expression profile in TCGA breast cancer patients. Sci Rep. 2021;11:1679.

31. Emmanuel N, Lofgren KA, Peterson EA, Meier DR, Jung EH, Kenny PA. Mutant GATA3 Actively Promotes the Growth of Normal and Malignant Mammary Cells. Anticancer Res. 2018;38:4435–41.

32. Kader T, Hill P, Zethoven M, Goode DL, Elder K, Thio N, et al. Atypical ductal hyperplasia is a multipotent precursor of breast carcinoma. J Pathol. 2019;248:326–38.

33. Kader T, Elder K, Zethoven M, Semple T, Hill P, Goode DL, et al. The genetic architecture of breast papillary lesions as a predictor of progression to carcinoma. NPJ Breast Cancer. 2020;6:9.

34. Cai Y, Crowther J, Pastor T, Abbasi Asbagh L, Baietti MF, De Troyer M, et al. Loss of Chromosome 8p Governs Tumor Progression and Drug Response by Altering Lipid Metabolism. Cancer Cell. 2016;29:751–66.

35. Thompson E, Taube JM, Elwood H, Sharma R, Meeker A, Warzecha HN, et al. The immune microenvironment of breast ductal carcinoma in situ. Mod Pathol. 2016;29:249–58.

36. Danforth DN Jr. Genomic Changes in Normal Breast Tissue in Women at Normal Risk or at High Risk for Breast Cancer. Breast Cancer . 2016;10:109–46.

37. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.

38. Risom T, Glass DR, Liu CC, Rivero-Gutiérrez B, Baranski A, McCaffrey EF, et al. Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma [Internet]. bioRxiv. 2021 [cited 2021 Apr 28]. p. 2021.01.05.425362. Available from: https://www.biorxiv.org/content/10.1101/2021.01.05.425362v1.full-text

39. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344:1396–401.

40. Roden DL, Baker LA, Elsworth B, Chan C-L, Harvey K, Deng N, et al. Single cell transcriptomics reveals molecular subtype and functional heterogeneity in models of breast cancer [Internet]. bioRxiv. 2018 [cited 2021 Apr 28]. p. 282079. Available from: https://www.biorxiv.org/content/10.1101/282079v1.full

41. Allred DC, Wu Y, Mao S, Nagtegaal ID, Lee S, Perou CM, et al. Ductal carcinoma in situ and the emergence of diversity during breast cancer evolution. Clin Cancer Res. 2008;14:370–8.

42. Sun R, Hu Z, Curtis C. Big Bang Tumor Growth and Clonal Evolution. Cold Spring Harb

Perspect Med [Internet]. 2018;8. Available from: http://dx.doi.org/10.1101/cshperspect.a028381

43. Polyak K. Is breast tumor progression really linear? Clin. Cancer Res. 2008. p. 339–41.

44. Zeng Z, Vo A, Li X, Shidfar A, Saldana P, Blanco L, et al. Somatic genetic aberrations in benign breast disease and the risk of subsequent breast cancer. NPJ Breast Cancer. 2020;6:24.

45. Silverstein MJ. The University of Southern California/Van Nuys prognostic index for ductal carcinoma in situ of the breast. Am J Surg. 2003;186:337–43.

46. Mannu GS, Wang Z, Broggio J, Charman J, Cheung S, Kearins O, et al. Invasive breast cancer and breast cancer mortality after ductal carcinoma in situ in women attending for breast screening in England, 1988-2014: population based observational cohort study. BMJ. 2020;369:m1570.

47. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst. 1989;81:1879–86.

48. Lee A, Mavaddat N, Wilcox AN, Cunningham AP, Carver T, Hartley S, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genet Med. 2019;21:1708–18.

49. Gil Del Alcazar CR, Huh SJ, Ekram MB, Trinh A, Liu LL, Beca F, et al. Immune Escape in Breast Cancer During In Situ to Invasive Carcinoma Transition. Cancer Discov. 2017;7:1098–115.

50. Kos Z, Roblin E, Kim RS, Michiels S, Gallas BD, Chen W, et al. Pitfalls in assessing stromal tumor infiltrating lymphocytes (sTILs) in breast cancer. NPJ Breast Cancer. 2020;6:17.

51. Hendry S, Salgado R, Gevaert T, Russell PA, John T, Thapa B, et al. Assessing Tumor-infiltrating Lymphocytes in Solid Tumors: A Practical Review for Pathologists and Proposal for a Standardized Method From the International Immunooncology Biomarkers Working Group: Part 1: Assessing the Host Immune Response, TILs in Invasive Breast Carcinoma and Ductal Carcinoma In Situ, Metastatic Tumor Deposits and Areas for Further Research. Adv Anat Pathol. 2017;24:235–51.

52. Cornel AM, Mimpen IL, Nierkens S. MHC Class I Downregulation in Cancer: Underlying Mechanisms and Potential Targets for Cancer Immunotherapy. Cancers [Internet]. 2020;12. Available from: http://dx.doi.org/10.3390/cancers12071760

53. Garrido MA, Rodriguez T, Zinchenko S, Maleno I, Ruiz-Cabello F, Concha Á, et al. HLA class I alterations in breast carcinoma are associated with a high frequency of the loss of heterozygosity at chromosomes 6 and 15. Immunogenetics. 2018;70:647–59.

54. Campbell MJ, McCune E, Bolen J, VandenBerg S, Chien J, Wong J, et al. Abstract 961: Intralesional injection of anti-PD-1 (pembrolizumab) results in increased T cell infiltrate in high risk DCIS. Cancer Res. American Association for Cancer Research; 2018;78:961–961.

55. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast

universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

56. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.

57. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177:1888–902.e21.

58. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20:296.

59. Gendoo DMA, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics. 2016;32:1097–9.

60. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.

61. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009;462:108–12.

62. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018;46:D649–55.

63. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27:1739–40.

64. Didion JP, Martin M, Collins FS. Atropos: specific, sensitive, and speedy trimming of sequencing reads. PeerJ. 2017;5:e3720.

65. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. arXiv [q-bio.GN]. 2013. Available from: http://arxiv.org/abs/1303.3997

66. Tischler G, Leonard S. biobambam: tools for read pair collation based algorithms on BAM files. Source Code Biol Med. 2014;9:13.

67. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

68. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44:e108.

69. Van der Auwera GA, O'Connor BD. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. "O'Reilly Media, Inc."; 2020.

70. Guimera RV. bcbio-nextgen: Automated, distributed next-gen sequencing pipeline. EMBnet.journal. 2011;17:30.

71. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly . 2012;6:80–92.

72. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.

73. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.

74. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46:D1062–7.

75. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.

76. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol. 2016;12:e1004873.

77. Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A. 2010;107:16910–5.

78. Nilsen G, Liestøl K, Van Loo P, Moen Vollan HK, Eide MB, Rueda OM, et al. Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. BMC Genomics. 2012;13:591.

79. Islam SMA, Ashiqul Islam SM, Wu Y, Díaz-Gay M, Bergstrom EN, He Y, et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor [Internet]. Available from: http://dx.doi.org/10.1101/2020.12.13.422570

80. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.

81. Kim S, Kim S, Kim J, Kim B, Kim SI, Kim MA, et al. Evaluating Tumor Evolution via Genomic Profiling of Individual Tumor Spheroids in a Malignant Ascites from a Patient with Ovarian Cancer Using a Laser-aided Cell Isolation Technique [Internet]. Available from: http://dx.doi.org/10.1101/282277

82. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27:592–3.

83. Reiter JG, Makohon-Moore AP, Gerold JM, Bozic I, Chatterjee K, Iacobuzio-Donahue CA, et al. Reconstructing metastatic seeding patterns of human cancers. Nat Commun. 2017;8:14114.

84. Mori H, Bolen J, Schuetter L, Massion P, Hoyt CC, VandenBerg S, et al. Characterizing the Tumor Immune Microenvironment with Tyramide-Based Multiplex Immunofluorescence. J Mammary Gland Biol Neoplasia. 2020;25:417–32.

85. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al.

QuPath: Open source software for digital pathology image analysis. Sci Rep. 2017;7:16878.

Figures & Tables

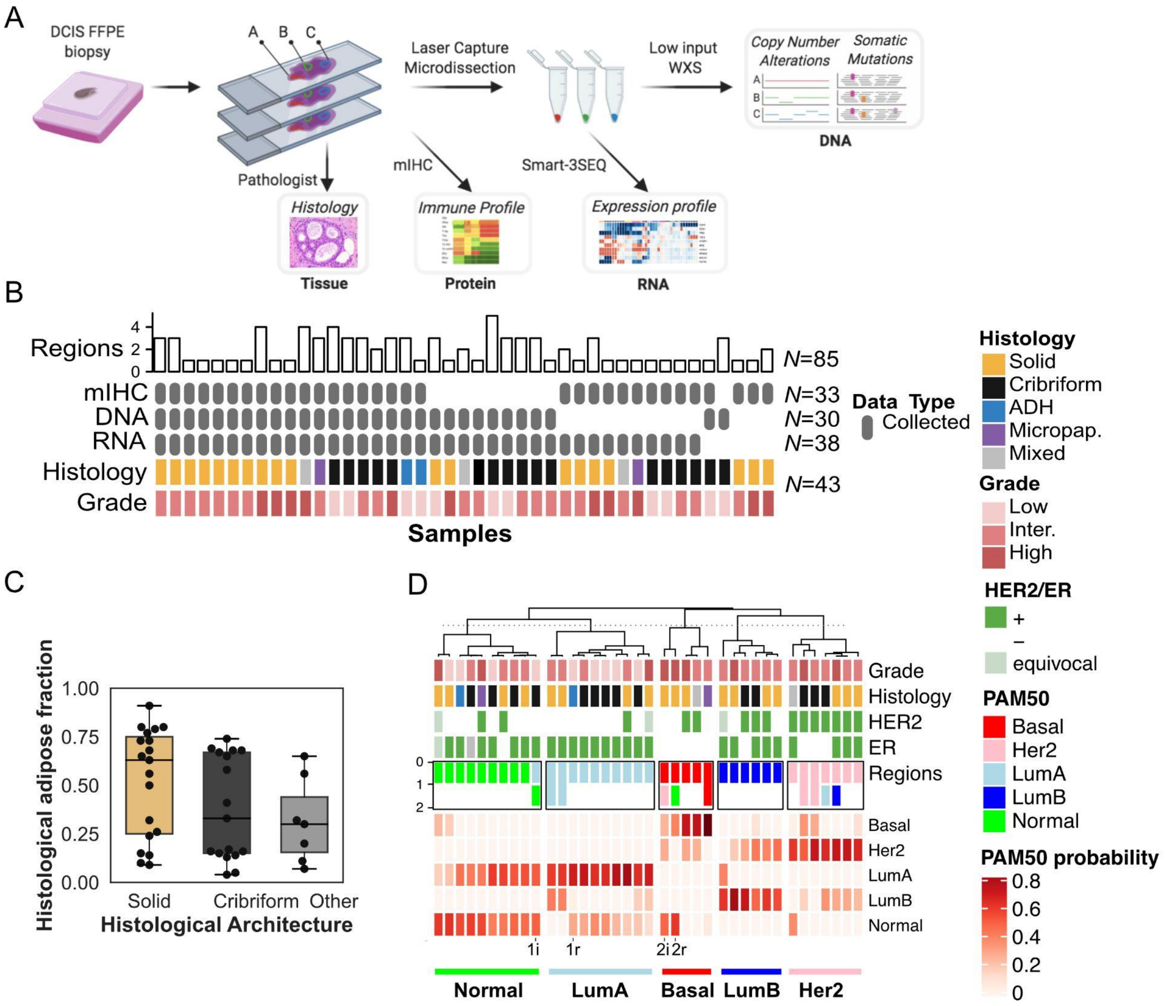


Figure 1: Study design and cohort overview (A) Archival sample processing and analysis workflow including histology (H&E and mIHC) and microbiopsy-derived whole exome (WXS) and whole transcriptome (Smart-3SEQ) profiling. **(B)** Study cohort overview including histological characteristics (colored rows), data type (grey rows) and number of histological regions (bar chart) investigated. **(C)** Estimate of the fraction of adipose area in H&E images in epithelium of the mIHC images according to each histological architecture. **(D)** Sample classification according to the probabilities of each PAM50 expression subtype. For 10 eligible samples, the intrinsic subtype of a spatially distinct region is indicated. Two patients with recurrence (r) and index (i) samples are indicated at the bottom.

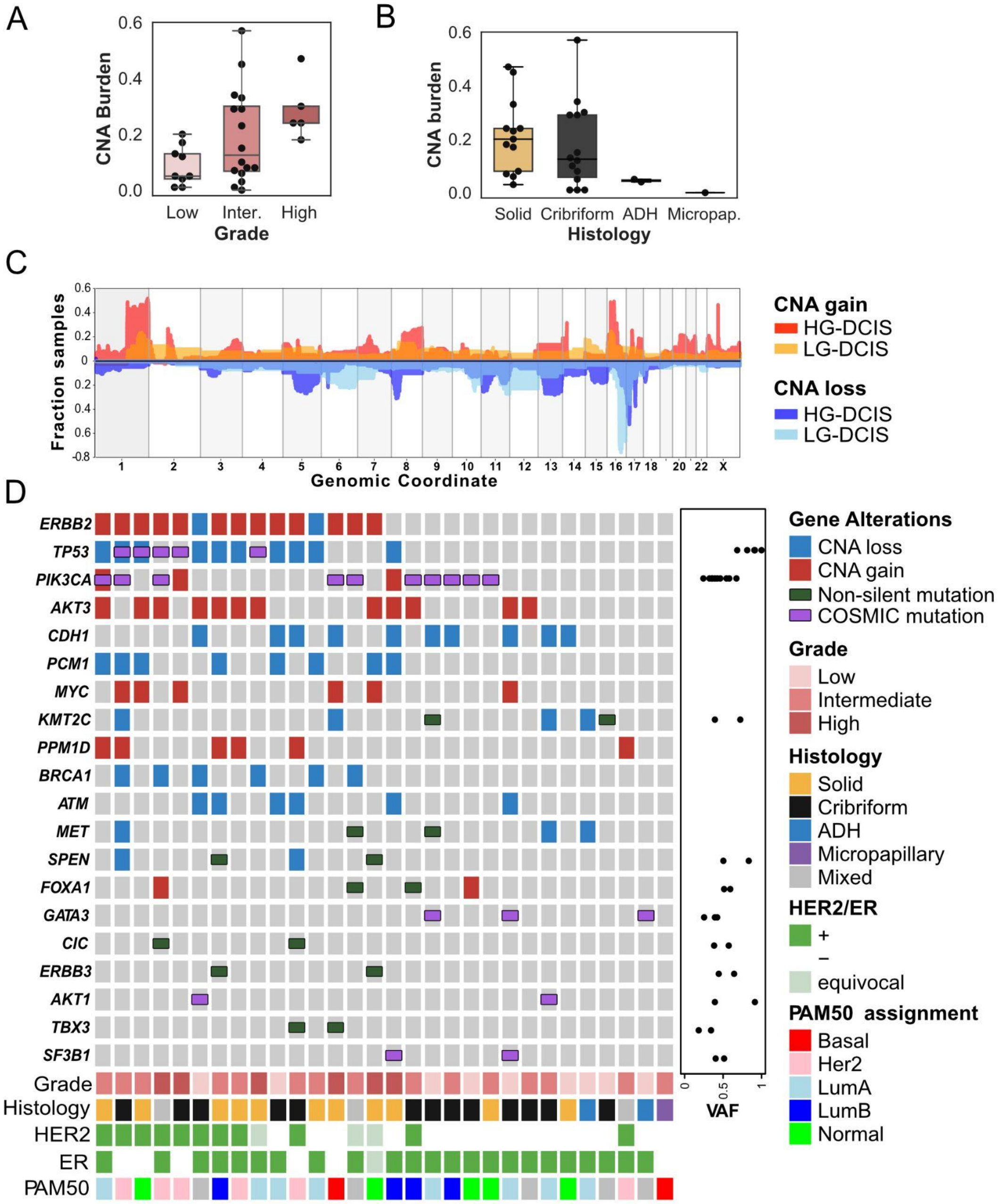
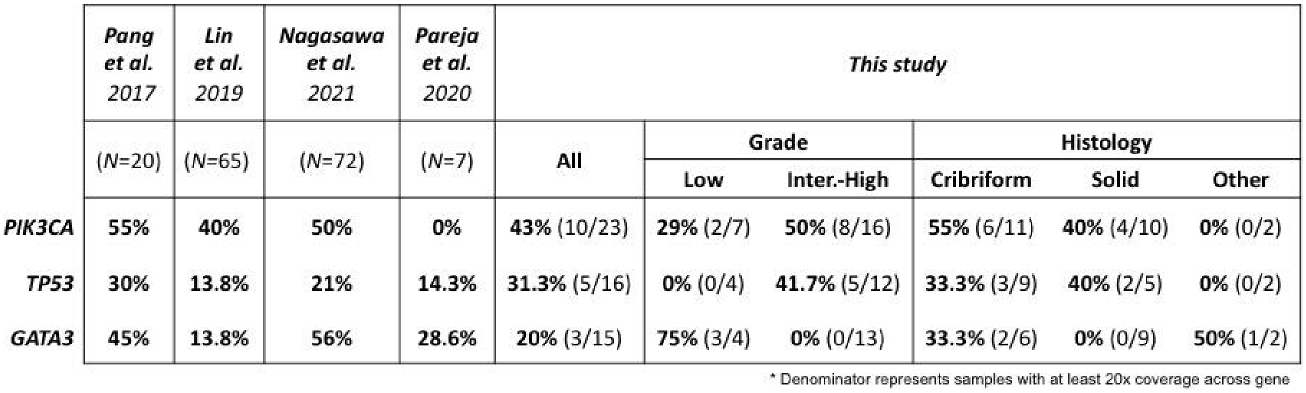


Figure 2. Pure DCIS genomic landscape (A-B) CNA burden (fraction of base pairs involved in copy number gain or loss) as a function of grade (A) and (B) histological architecture. **(C)** Smoothed frequency (y-axis) of CNA gains (top) and losses (bottom) smoothed along the genome (x-axis) for HG-DCIS (N=22 - dark colors) and LG-DCIS (N=5 light colors). **(D)** Oncoprint diagram displaying the mutational status of driver genes commonly altered in breast cancer. Genes were included if they were mutated in at least 2 patients or located in a CNA segment present in at least 6 patients, and ordered by frequency of alteration. The variant allele fraction (VAF) of mutations (right panel) and histological characteristics (bottom panel) are indicated.

Table 1. Frequency of *PIK3CA*, *TP53* and *GATA3* driver mutations in previously reported (Pang et al., Lin et al., Nagasawa et al., Pareja et al.) and current pure DCIS cohort.



	Pang et al. 2017	*Lin et al. 2019*	*Nagasawa et al. 2021*	*Pareja et al. 2020*	*This study*					
	(*N*=20)	(*N*=65)	(*N*=72)	(*N*=7)	**All**	**Grade**		**Histology**		
						Low	**Inter.-High**	**Cribriform**	**Solid**	**Other**
PIK3CA	**55%**	**40%**	**50%**	**0%**	**43%** (10/23)	**29%** (2/7)	**50%** (8/16)	**55%** (6/11)	**40%** (4/10)	**0%** (0/2)
TP53	**30%**	**13.8%**	**21%**	**14.3%**	**31.3%** (5/16)	**0%** (0/4)	**41.7%** (5/12)	**33.3%** (3/9)	**40%** (2/5)	**0%** (0/2)
GATA3	**45%**	**13.8%**	**56%**	**28.6%**	**20%** (3/15)	**75%** (3/4)	**0%** (0/13)	**33.3%** (2/6)	**0%** (0/9)	**50%** (1/2)

* Denominator represents samples with at least 20x coverage across gene

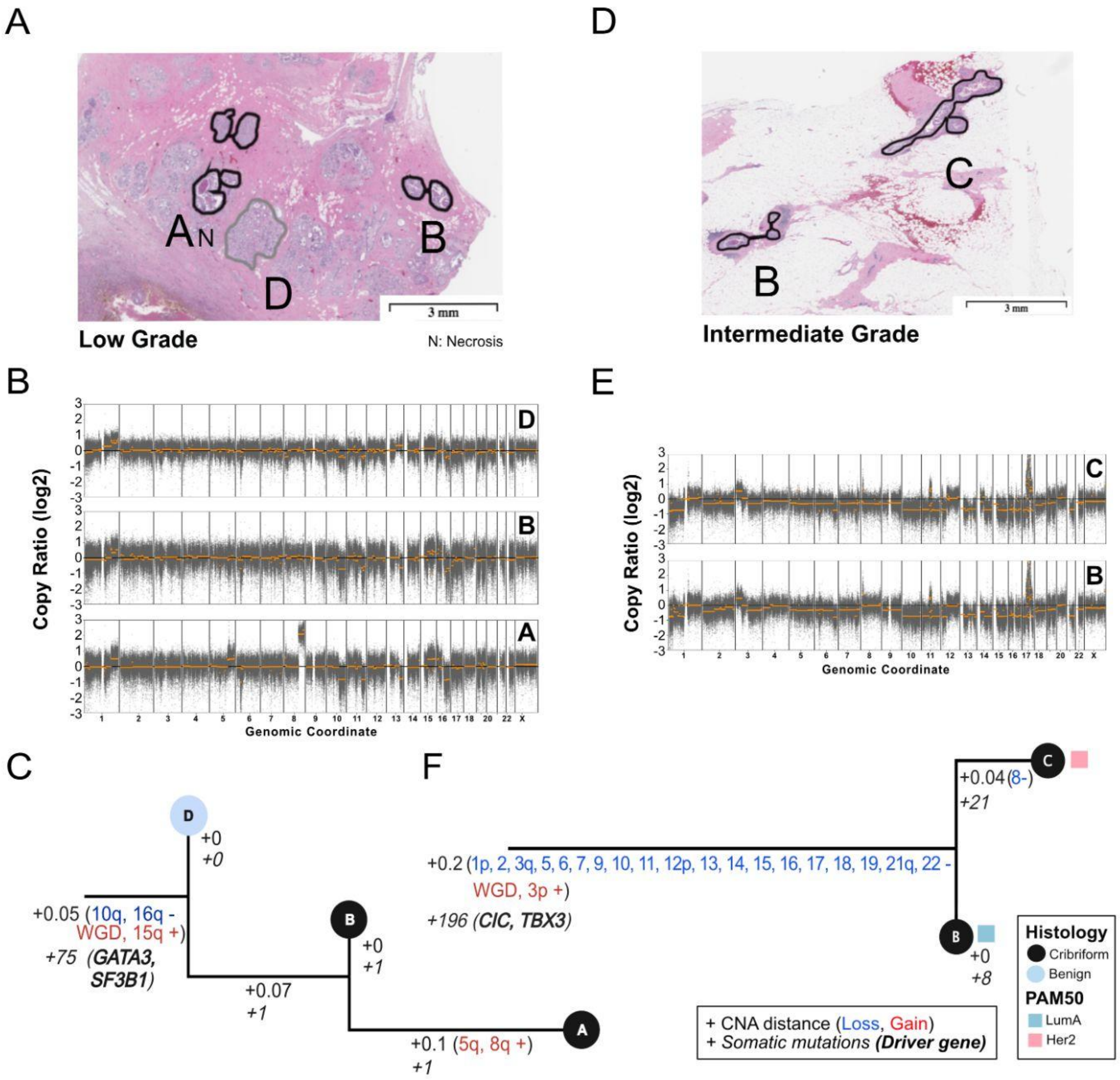


Figure 3. Clonal relationships of multi-region DCIS. Multi-region phylogenetic reconstruction using both CNA and somatic mutations for MCL76_061_16200 **(A-C)** and MCL76_077_15300 **(D-F)**. For each case, the spatial annotation of the microdissected regions on the H&E images (A,D), corresponding copy number profiles (B,E) and phylogenetic trees (C,F) are displayed. Copy number profile plots show bins (grey dots) and segment (orange) $\log_2$ copy number ratio (y-axis). The phylogenetic tree leaves (single dissected region) are colored according to histological type and the branches (hamming distances based on CNA segments) annotated with corresponding specific somatic alterations or their total number (CNA: regular, genes: italic font). The tree root corresponds to an inferred normal diploid ancestor. PAM50 subtype of the region is indicated when available. Annotations and trees are available for 10 additional samples in Figure S4.

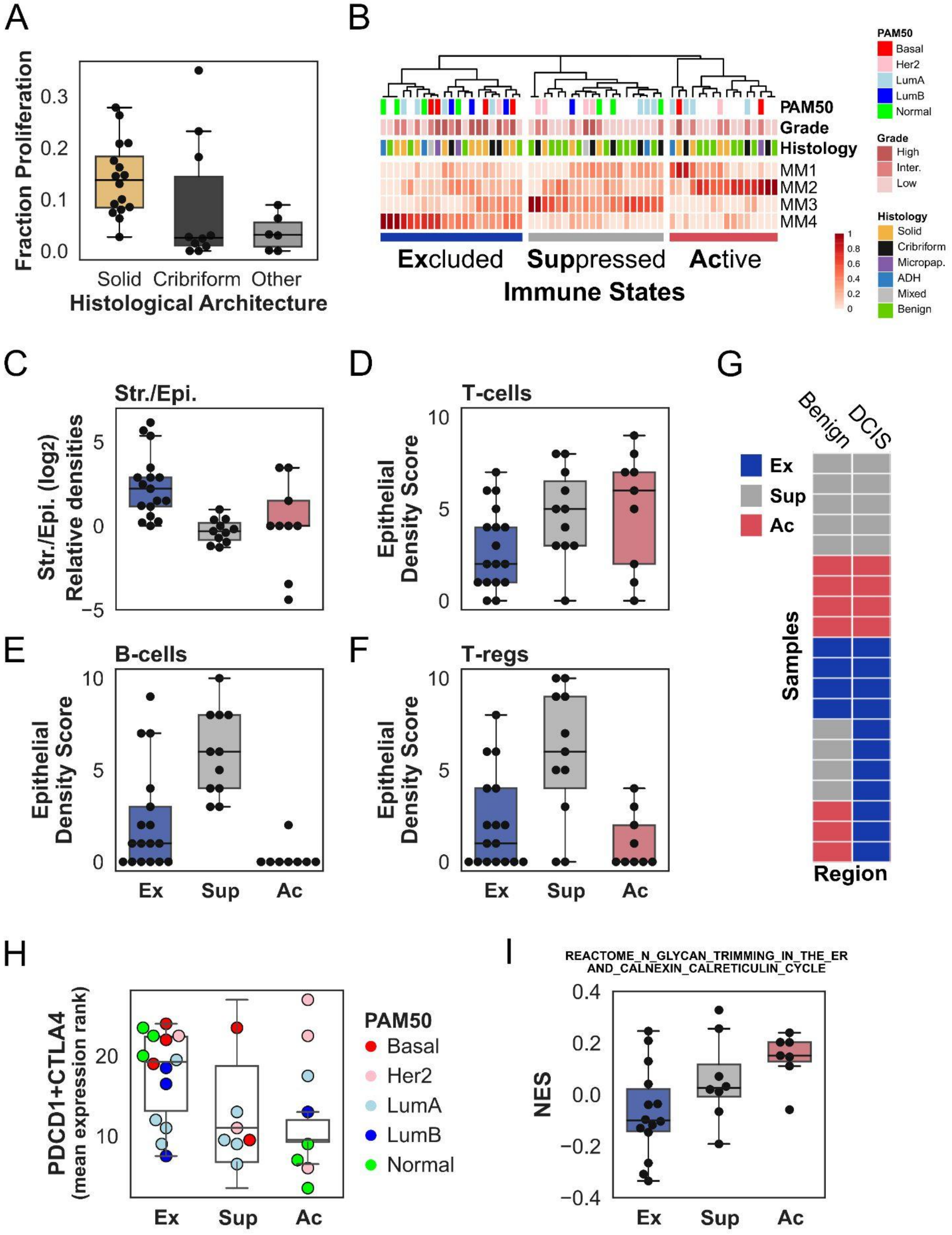


Figure 4. Characterization of the immune landscape. (A). Estimates of the fraction of Ki67+ cytokeratin positive cells in epithelium of the mIHC images according to each histological architecture. **(B)** Classification of benign and premalignant regions into 3 immune states according to Meta-Marker (MM) values derived from a non-negative matrix factorization

(described in Figure S5). Each state is named after Meta-Marker composition: Active (Ac: MM1&MM2 high TC densities), Suppressed (Sup: MM3 high BC and TREG densities) or Excluded (Ex: MM4 high - low epithelial densities). **(C-F)** Distribution of relative stromal total immune-cell density (C), epithelial cell density scores for T-cells (D), B-cells (E), or T-reg (F) between the 3 immune states. Density scores are obtained from median decile of the cell densities. **(G)** Immune-state comparison in 20 samples (rows) with matching benign (left column) and premalignant (right column) regions. **(H)** Expression of immune checkpoint receptors genes, *PDCD1* and *CTLA-4* in each immune state. **(I)** GSEA normalized enrichment score (NES) for a Reactome gene set across immune states.

Supplemental Figures:

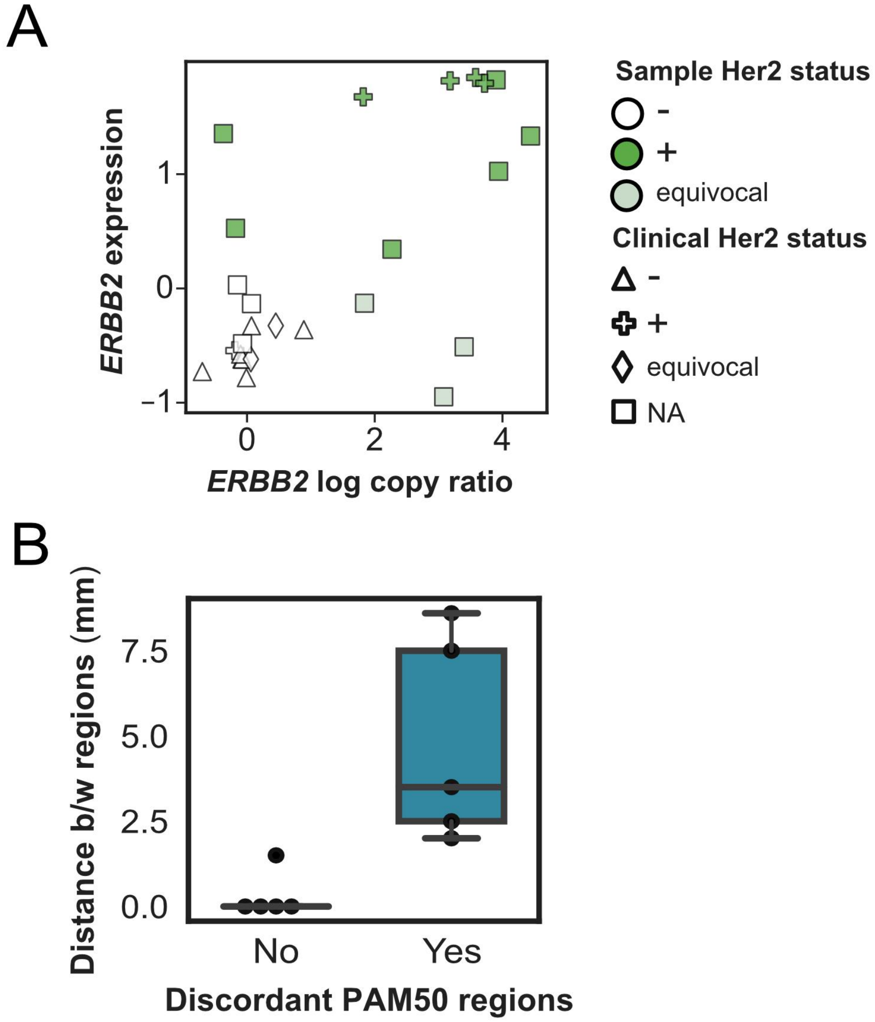


Figure S1: Pure DCIS characterization (A) Estimation of Her2 status. The DNA-based $\log_2$ copy number ratio (x-axis) and RNA-based expression level (y-axis) of *ERBB2* gene are displayed for 26 samples with both data available. **(B)** PAM50 discordance between regions in relationship with spatial distance between regions.

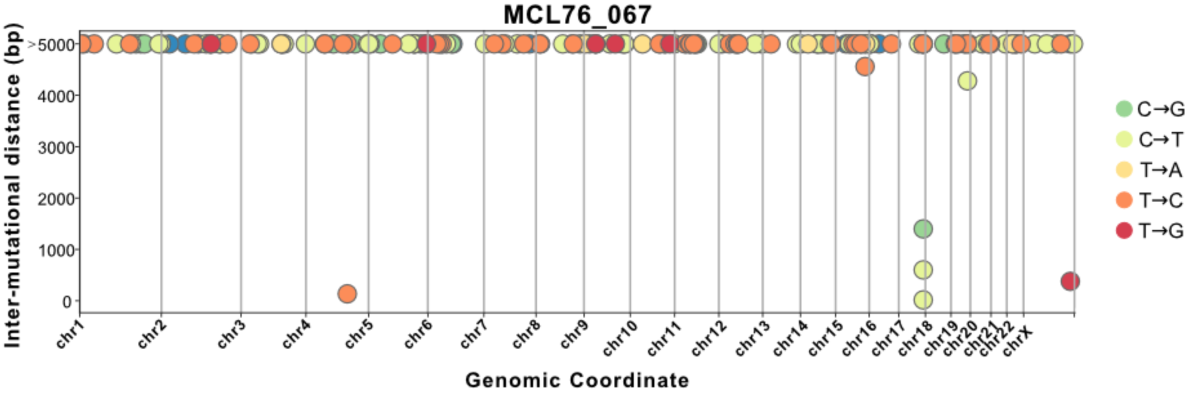


Figure S2. Likely kataegis event in MCL76_067_16600 in chromosome 17. Along the genome coordinate (x-axis), the relative distance between proximal mutations (y-axis) as well as their substitution type (colors) are indicated.

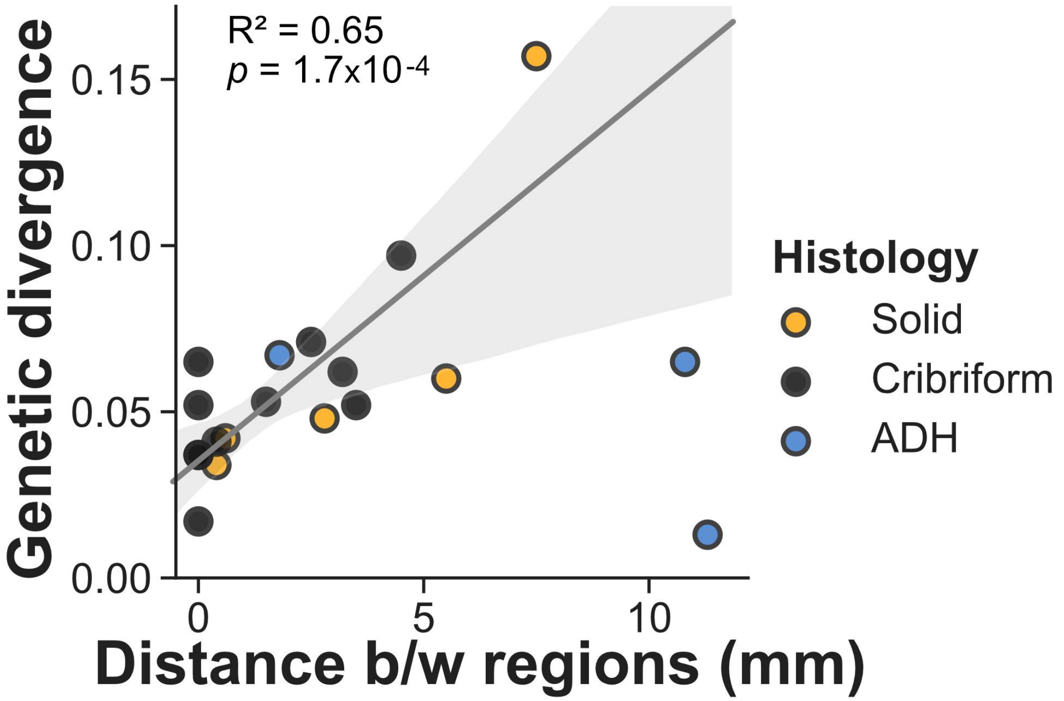
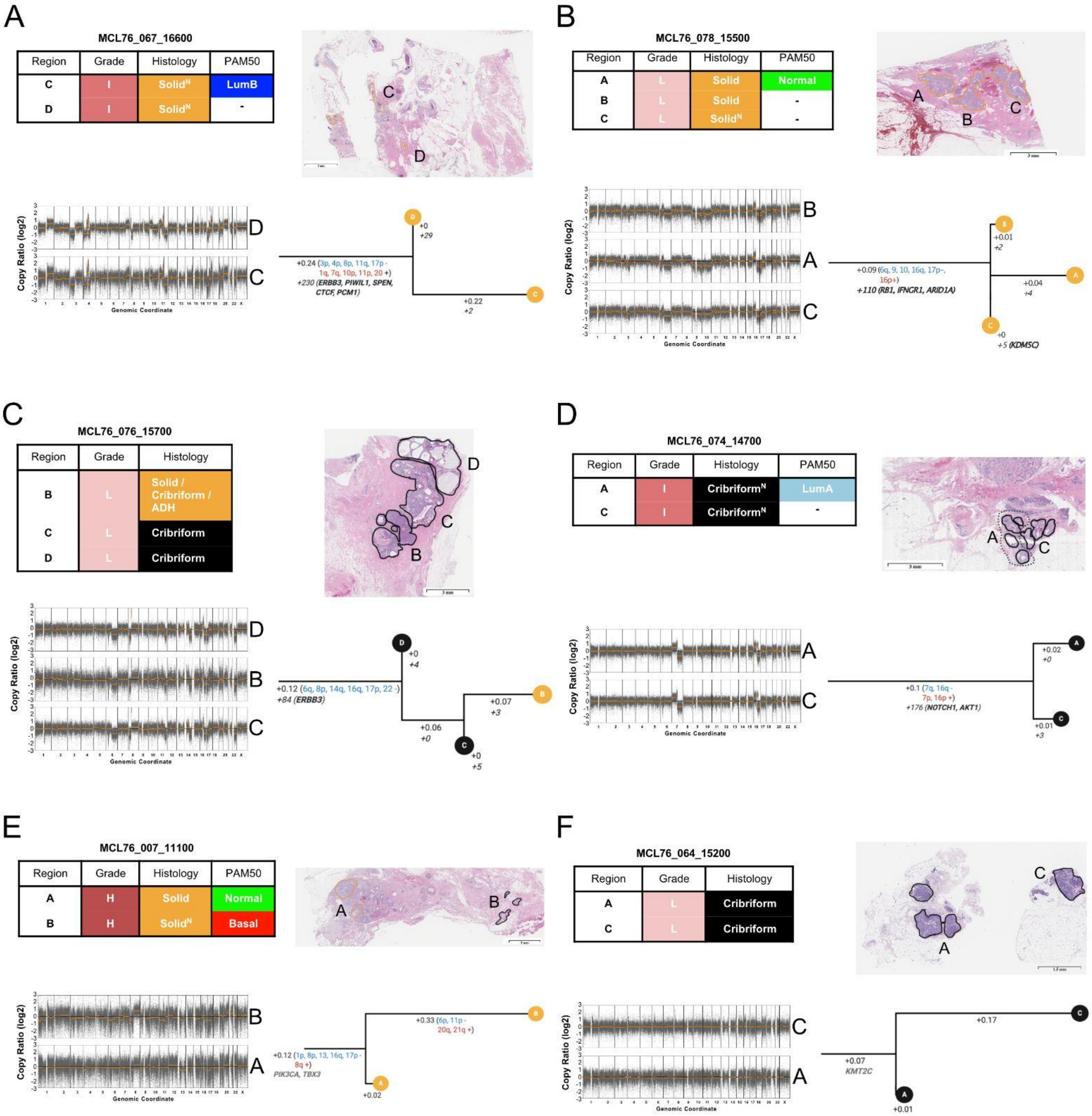


Figure S3. Genetic divergence in multi-region DCIS. CNA-based genetic divergence (y-axis) of each pair of histologically concordant regions (dot) as a function of the minimum physical distance between them (x-axis), colored by their histology. Linear regression line fit of DCIS samples shown in solid dark gray line, with 95% confidence interval estimate based on bootstrapping in light gray.



A

MCL76_067_16600

Region	Grade	Histology	PAM50
C	I	Solid^{N}	LumB
D	I	Solid^{N}	-

B

MCL76_078_15500

Region	Grade	Histology	PAM50
A	L	Solid	Normal
B	L	Solid	-
C	L	Solid^{N}	-

C

MCL76_076_15700

Region	Grade	Histology
B	L	Solid / Cribriform / ADH
C	L	Cribriform
D	L	Cribriform

D

MCL76_074_14700

Region	Grade	Histology	PAM50
A	I	Cribriform^{N}	LumA
C	I	Cribriform^{N}	-

E

MCL76_007_11100

Region	Grade	Histology	PAM50
A	H	Solid	Normal
B	H	Solid^{N}	Basal

F

MCL76_064_15200

Region	Grade	Histology
A	L	Cribriform
C	L	Cribriform

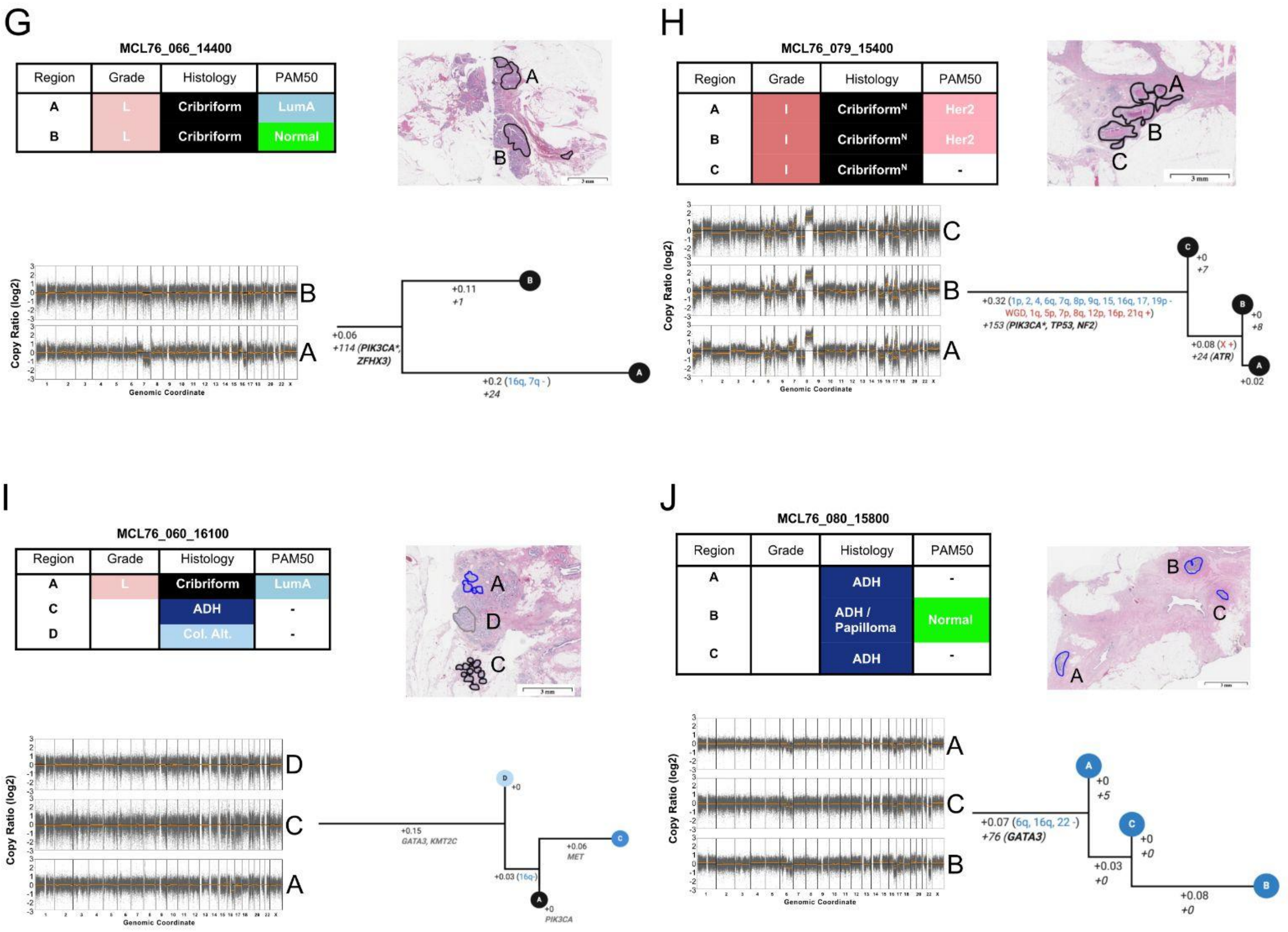


Figure S4. Phylogenetic trees for multi-region DCIS samples. (A-J) Clonal reconstruction using CNA and somatic mutations in 25 related regions across 10 samples. In each panel, _top left:_ a table describing the name, nuclear grade and histological architecture of each region in a sample is shown, (necrosis is indicated with N); _top right:_ shows an H&E image of the sample with dissected regions drawn on the image; _bottom left_: Copy number profiles for each region in the sample, genomic coordinates are indicated on the X-axis and the $\log_2$ copy ratio on the Y-axis. Bins are indicated in dark-grey and segments in orange; _bottom right_: Phylogenetic tree for the sample with leaf nodes indicating a single dissected region colored by histology. The tree is rooted to a normal diploid ancestor. Branch lengths are hamming distances based on CNA segments. Branches are labeled with 1) CNA-based branch length, and, when available, 2) arm-level CNA losses (blue) and gains (red) and 3) somatic mutation number (italic) with mutations in breast cancer driver genes indicated (black: high coverage, grey: low coverage). CNA smaller than arms, or on driver genes are not displayed.

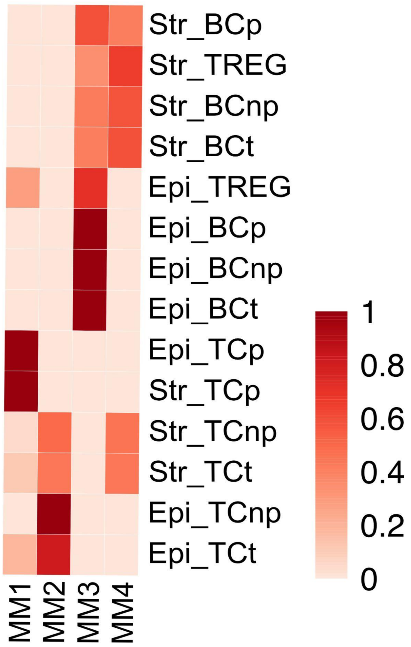


Figure S5. Composition of the NMF Meta-markers (columns MM1-4) according to the densities scores (red scale) of each cell type (BC: B-cells, TC: T-cells, TREG: regulatory T-cells), proliferative state (p: Ki67+, np: KI67-, t: total) and regional location (Epi: Epithelium, Str: Stroma).Estimates of the fraction of Ki67+ cytokeratin positive cells in epithelium of the mIHC images according to each histological architecture.

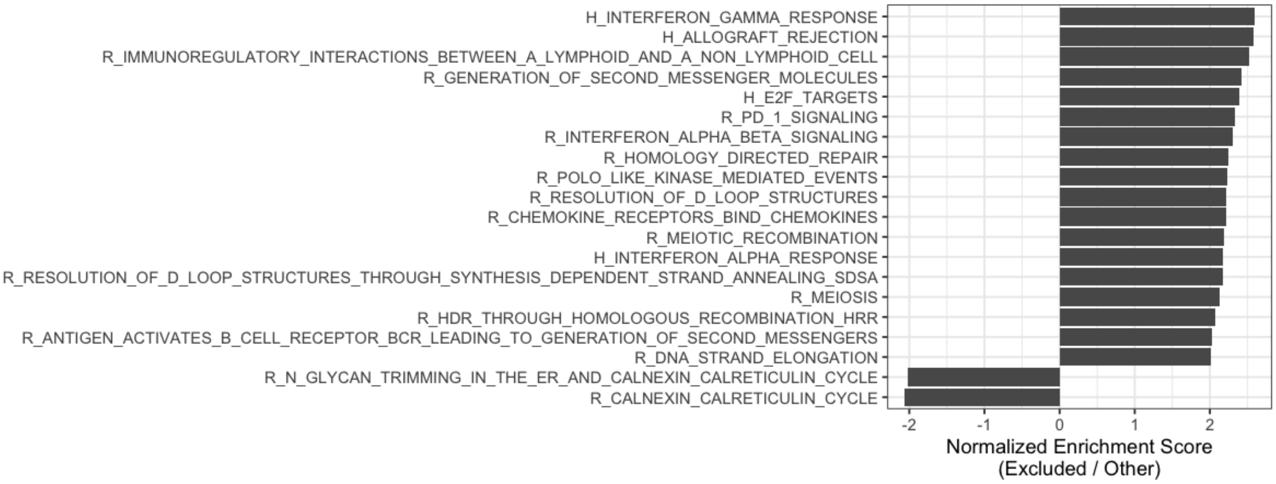


Figure S6: Gene sets significantly deregulated in epithelium of regions in the Excluded immune state. All Hallmark (H) and Reactome (R) genesets were tested. Genesets with an absolute normalized enrichment score greater than 2 and with FDR less than 0.05 are represented.