Eco-evolutionary Guided Pathomics Analysis to Predict DCIS Upstaging

Yujie Xiao[1], Manal Elmasry[2,3], Ji Dong K. Bai[2], Andrew Chen[2], Yuzhu Chen[2], Brooke Jackson[4], Joseph O. Johnson[4], Robert J. Gillies[4], Prateek Prasanna[5], Chao Chen[5*], Mehdi Damaghi[1,2,3*]

1- Department of Applied Mathematics and Statistics, Stony Brook University, NY, USA
2- Department of Pathology, Stony Brook Medicine, Stony Brook University, NY, USA
3- Department of Pathology, Faculty of Medicine, Mansoura University, Mansoura, Egypt
4- Moffitt Cancer Center, Tampa, Fl, USA
5-Department of Biomedical Informatics, Stony Brook Medicine, Stony Brook University, NY, USA

Running title: Eco-Evolutionary Designed Biomarker

* Corresponding authors: Chao Chen: chao.chen.1@stonybrook.edu,
Mehdi Damaghi: mehdi.damaghi@stonybrookmedicine.edu
Stony Brook Cancer Center
MART building, 9th floor, 9M0802
Lauterbur Drive
Stony Brook, NY 11794-7263
Phone: (631) 216-2920

Declaration of interests

The authors declare no competing interest.

Abstract

Cancers evolve in a dynamic ecosystem. Thus, characterizing cancer's ecological dynamics is crucial to understanding cancer evolution and can lead to discovering novel biomarkers to predict disease progression. Ductal carcinoma in situ (DCIS) is an early-stage breast cancer characterized by abnormal epithelial cell growth confined within the milk ducts. In this study, we show that ecological habitat analysis of hypoxia and acidosis biomarkers can significantly improve prediction of DCIS upstaging. First, we developed a novel eco-evolutionary designed approach to define habitats in the tumor intra-ductal microenvironment based on oxygen diffusion distance. Then, we identified cancer cells with metabolic phenotypes attributed to their habitat conditions, such as the expression of CA9 indicating hypoxia responding phenotype, and LAMP2b indicating the acid adaptation. Traditionally these markers have shown limited predictive capabilities for DCIS upstaging, if any. However, when analyzed from an ecological perspective, their power to differentiate between pure DCIS and upstaged DCIS increased significantly. Second, using eco-evolutionary guided computational and digital pathology techniques, we discovered distinct niches with spatial patterns of these biomarkers and used the distribution of such niches to predict patient upstaging. The niches patterns were characterized by pattern analysis of both cellular and spatial features. With a 5-fold validation on the biopsy cohort, we trained a random forest classifier to achieve the area under curve (AUC) of 0.74. Our results affirm the importance of using eco-evolutionary-designed approaches in biomarkers discovery studies in the era of digital pathology by demonstrating the role of tumor ecological habitats and niches.

Keywords:

Tumor ecology and evolution, DCIS, Eco-evolutionary biomarkers, Metabolic phenotypes, Habitat analysis, Niche analysis, Pathomics, Machine learning, Digital pathology

Introduction:

In recent years, the understanding that cancer is a dynamic ecological and evolutionary process has become deeply entrenched (1,2,3). To date, several evolutionary approaches have been adapted and applied in cancer biology, such as diversity measures to predict disease progression; however, tumor ecosystem and ecological habitat and niche studies are still overlooked (3,4). Within the human body and much like organisms in the natural world, cancer cells follow evolutionary principles, utilizing resources and establishing habitats and niches within tissues (5,6). This ecological perspective of cancer is crucial for discovering the natural selection driving cancer evolution. Recognizing the parallels between organismal ecology and the tumor microenvironment opens up untapped opportunities to incorporate ecological measures, improving our understanding of both tumor dynamics and selective pressures shaping tumors' evolutionary landscapes. Such insights may potentially lead to improved cancer prognosis, progression prediction, risk stratification, and therapeutic strategies. If tumor evolutionary state and/or its evolutionary trajectories could be reliably achieved using a single biopsy tissue, clinical translation would be comparatively more manageable. Nevertheless, studies have yet to determine whether measures of tumor evolvability derived from a single biopsy sample are adequate, or if the inclusion of multiple samples significantly enhances predictions of clinical outcomes (7).

Breast cancer incidence in the US has been increasing over the past decade at a rate of 0.5% per year(8). With increased mammographic screening, there has been a substantial increase in detecting the early non-invasive forms of breast cancer, such as ductal carcinoma in situ (DCIS)(2,9). About one-third of breast cancers detected by mammography are DCIS (10). As the most common pre-cancer state, DCIS can progress to invasive disease in a linear evolution pattern, or can be part of other clonal evolutionary dynamics such as branching, punctuated, or neutral evolution (2,9,11). Since DCIS and IDC (invasive ductal carcinoma) are indistinguishable by (epi-)genetic mutations, gene expression, or protein biomarkers, and because it is not possible to predict whether DCIS will remain indolent or progress to more aggressive disease, almost all early tumors are treated with aggressive interventions(2,12–14). To avoid such over treatment, more research is needed to fully understand evolution from pre-cancer to indolent DCIS or progress to IDC(9).

DCIS is a heterogeneous group of neoplastic lesions confined to the mammary ducts. The confinement of proliferating neoplastic cells inside the duct and growth of pre-cancer cells toward the center of the duct, which is far from vasculature, causes limitations in oxygen and nutrients. This intraductal oxygen microenvironment is also influenced by complex ecosystems surrounding the duct, such as vascular activity(15), stiffness of extracellular matrix (ECM) (16), and metabolites (6,17,18,19) (**Figure 1A**) . Local microinvasion is the main difference between DCIS and IDC and might also be the first evolutionary step of progressing in the case of linear evolution(11). Microinvasion consists of cohorts of cancer cells that breach the basement membrane into the surrounding ECM. Recently, genomic analysis of matched DCIS and IDC samples has revealed that in 75% of cases, the invasive recurrence was found to be clonally related to the initial DCIS. This implies that tumor cells derived from DCIS could evolve in a linear or branching fashion with 18% new transformations and/or clonogenesis (11). These new findings emphasize the extraordinary heterogeneity in genotype and phenotypic plasticity in breast cancer that must be studied in the light of evolution and ecological studies. Thus, we designed our study to capture the phenotypic heterogeneity of cancer cells in their selective microenvironments. We hypothesize that non-genetic ecological factors, such as intra-ductal microenvironmental conditions, may be responsible for transitioning from a DCIS to IDC phenotype, in the case of linear and branching evolution, or may select clones with pre-existing IDC phenotypes in the case of the other evolutionary trajectories, including punctuated and neutral evolution(6,11,18,20).

To validate this hypothesis, we propose a novel method to study DCIS evolution, by capturing and characterizing "tumor habitats" and "cell niches" and their interactions in the tumor ecosystem. Natural selection requires phenotypic diversity within a population undergoing microenvironmental selection forces (21). Cells that adapt in response to natural selection may present similar phenotypes, corresponding to the microenvironment exerting the selection. We started by defining the habitats based on availability of oxygen into: a) oxygenated habitat and b) hypoxic habitat. Following previous theory(18,22), these habitats are defined by distance from the duct boundary. However, a uniform distance threshold hardly captures the true oxidate/hypoxic states of cells. Therefore, we further proposed to fine-tune these habitats using protein expression indicative of phenotypes resulting from cancer cell adaptation to variation in oxygen availability. Therefore, we defined *intraductal DCIS niches* inside habitats as clusters of cells with similar phenotypic behavior responding to hypoxia. Through analysis via these niches, we can identify more aggressive phenotypes leading to microinvasion and DCIS upstaging to IDC or possible direct evolution to IDC without going through DCIS sub-stages.

Our biomarkers are designed based on prior biological knowledge. Oxygen availability determines the source of energy production as of either mitochondrial respiration or glycolysis. Hypoxic cells switch to glycolysis, causing lactic acid production that can lead to acidosis when lactic acid is not cleared from the tumor space. Peri-luminal cells will experience hypoxia if they are far (>0.125 - 0.160 mm) from a blood supply. These cancer cells inhabit a microenvironment of hypoxia, acidosis, and severe nutrient deprivation (18,22). These environmental properties exert a strong selection pressure upon the cancer cells, which in turn feeds back to the microenvironment, creating a dynamically changing tumor ecosystem containing several habitats. We have shown that cancer cells within breast ducts subjected to chronic hypoxia and acidosis evolve mechanisms of adaptations to survive in this harsh microenvironment (17,18,20). We have also shown that cells adapted to hypoxic and/or acidic niches have developed specific metabolic vulnerabilities that can be targeted to push them back to a more physiologically normal state(17). Both these studies strengthen the acid-induced evolution model of breast cancer and our proposed evolutionary designed biomarkers including CA9 and LAMP2b in this research(6,17,20,23,24). Here we examined the role of these biomarkers within an eco-evolutionary concept as a predictor of DCIS upstaging for the first time. We used these markers as representative of the cancer cell metabolic states to define niches inside habitats that can select for more aggressive phenotypes, leading to microinvasion and DCIS upstaging to IDC or possible direct evolution to IDC without going through DCIS sub-stages.

To perform our analysis, we curated a retrospective cohort of DCIS patients, with specimens collected from Biopsy (Bx) samples before surgery and after Excision (Ex). All the patients had histologically confirmed DCIS on core biopsy, followed by diagnosis confirmed on surgical excision specimens with either DCIS or IDC (**Figure 1B**). Our niche-based prediction model is trained and tested on the Bx samples. These best fits future clinical applications that machine learning model can be subsequently applied to predict upstaging at Bx for future patients. We then stained 3 sequentially sectioned slides for hematoxylin and eosin (HE), CA9 and LAMP2b. We manually annotated ducts bigger than 400 μms in diameter. The 200 μms in radius annotation ensures each duct has both oxygenated and hypoxic habitats to build a balanced cohort for analysis. We developed a novel algorithm to detect intra-ductal DCIS cell niches based on biomarker expression similarity. Then, we studied the spatial organization of CA9- and LAMP2b-positive cells as the eco-evolution markers of cancer cells in hypoxic and acidic habitats at three different scales: whole slide, duct, and oxygen habitats (normoxic and hypoxic). We also applied multiple spatial functions and spatial entropies were used to define niche and micro-niches describing the spatial patterns of the cell groups. After a systematic and comprehensive analysis, we observed that the spatial features at the finest habitat level possess the most predictive power where the micro-niches were defined by the expression of CA9 and LAMP2b in hypoxic habitats. By characterizing these niches and micro-niches with spatial and pathomics features, we then developed a risk scoring system by integrating principles of ecological-evolutionary dynamics with pathological imaging and molecular features of early-stage breast tumors (**Figure 1C**). We show that quantitative analyses of immuno-histological images combined with the tumor’s eco-evolution dynamics and underlying molecular pathophysiology can significantly improve predicting if the neoplasm has already evolved to invasive disease and cancer. We developed a machine learning model fine-tuning the tumor habitats into micro-niches using specific molecular signatures of resident cancer cells to provide informed decision support. In summary, we show that specific habitats containing micro-niches of cells with similar phenotypes responding to hypoxia and acidosis, or adaptation to long term exposure of these conditions, are responsible for DCIS progression, and hence would be correlated to upstaging. To test this hypothesis, we applied machine learning techniques to calculate the niches inside the tumor to define spatial and temporal distribution of habitats in solid tumors of DCIS patients with pure DCIS and upstaged disease. By deploying eco-evolutionary principles and machine learning techniques, our work proposes a novel consilient approach - as opposed to the traditional single biomarker studies - to stratify DCIS patients

Materials and Methods

Method overview

Our evolutionary analysis pipeline takes 3 consecutive slides of each patient sample, detects intra-ductal cell niches, characterizes these niches with their spatial and morphological features, and then predicts whether the patient will be pure DCIS or upstaged based on the distribution of these niches. In particular, the pipeline has 4 modules. First, we annotate and align ducts from different whole slide images (WSIs) of the same patient sample. This ensures cells of different slides are aligned and we can characterize their interactions. In the second module, we detect and map all eco-evo positive cells (i.e., cells activated with the selected stains) into the same duct and detect different clusters of cells as niches. In the third module, we characterize these niches with comprehensive spatial statistical features, as well as their morphological features as observed in HE. Finally, we categorize these niches into different subclasses through deep learning-based dimension reduction and clustering based on their features. We use the distribution of different niche subclasses to characterize different samples/patients. We demonstrate the discriminative power of this niche-based characterization in predicting whether a patient will be pure DCIS or upstaged in the future. **Figure 1C** illustrates the overview of our pipeline.

Data preparation and usage

The data used in this study is the biopsy samples collected after mammography and before surgery. 84 samples including 68 pure DCIS and 16 progressed to IDC were analyzed. This study complied with the Health Insurance Portability and Accountability Act and was approved by the institutional review board, with a waiver of the requirement for informed consent. Women with a core biopsy diagnosis of DCIS between 2012 and 2022 who consented to at Moffitt Cancer Center Total Cancer Care protocol were included in this analysis. Cases were excluded if surgical excision was performed more than 6 months after the core biopsy, if there was concurrent ipsilateral invasive breast cancer or metastatic malignancy, or if neoadjuvant chemotherapy (for a concurrent contralateral breast malignancy) or chemotherapy for a non-breast primary malignancy was administered between the dates of the DCIS core biopsy and surgery. Additional exclusions included a personal history of invasive breast cancer or DCIS within 12 months preceding the core biopsy or a concurrent diagnosis of Paget disease in the ipsilateral breast. After applying these inclusion and exclusion criteria, 84 cases of biopsy-proven DCIS were identified, of which 16 were upgraded at surgery and 68 remained non-upgraded.

Pure DCIS and upstaged patients were matched across clinical features, including age, race, ethnicity, grade, ER status, and PR status, to minimize their influence on the analysis (**Figure S1**). To validate the comparability of these groups, we conducted a Wilcoxon rank-sum test for the continuous variable (age) and chi-square tests for the categorical variables (race, ethnicity, grade, ER status, and PR status). None of these tests showed significant differences between the two groups, with all p-values larger than 0.1, indicating that the groups were well-matched.

For each sample, we obtained 3 whole slide images, including 1 HE and 2 IHC slides. We conducted 5-fold stratified cross validation, where 4 folds are used for niche clustering and for the training of the pure DCIS/upstaged classifier and 1-fold is used for validation. This fits the clinical application we are aiming for; we would like our model to estimate the risk based on biopsy samples, which are much less invasive and can be used for patient stratifications before surgery and hopefully decrease over treatment. Further details on HE and IHC acquisition are provided below.

Sample selection, immunohistochemistry and HE staining. Patients' tumor blocks were selected by pathologists using the archived HE stained slides. The blocks were sequentially sectioned 4 μms and de-identified for research use. 3 slides were stained with primary antibodies of 1:100 dilution of anti-LAMP2 (#ab18529, Abcam), and 1 ug/ml concentration of anti-CA9 (#AF2188, R&D), and HE staining using standard hematoxylin and eosin protocol. Positive and negative controls were used. Normal placenta was used as a positive control for LAMP2b and clear cell renal cell carcinoma was used as a positive control for CA9. For the negative control, an adjacent section of the same tissue was stained without application of primary antibody and any stain pattern observed was considered as non-specific binding of the secondary. Primary immunohistochemical analysis was conducted using digitally scanning slides. The scoring method used by the pathologist reviewer to determine (a) the degree of positivity scored the positivity of each sample ranged from 0 to 3 and was derived from the product of staining intensity (0–3+). A zero score was considered negative, score 1 was weak positive, score 2 was moderate positive, and score 3 was strong positive. (b) The percentage of positive tumors stained (on a scale of 0-3). Whole slide imaging (WSI) of IHC and HE slides were obtained by scanning at 20X magnification (of 0.5022 micrometer per pixel) using Aperio AT2 from Leica Biosystems. Images were transferred to cloud storage and locally to be uploaded in QuPath software for analysis. QuPath software was used to detect the positive pixels for each IHC marker (CA9 and LAMP2b) and to segment the HE images into hypoxic and normoxic tumor habitats based on their distance from the basement membrane. The 'Positive Cell Detection' function from Qupath was used to automatically classify the positivity of CA9 and LAMP2b markers and validated by the study pathologist.

MODULE 1: Duct annotation and alignment

Manual annotation of ducts in the Bx cohort. We annotate and align ducts within all input slides (1 HE + 2 IHCs per sample). After annotating ducts, we align the ducts from the three modalities via co-registration. This alignment enables us to map cells into the same spatial domain and analyze their interaction. Details are provided below. QuPath was used as the interface to annotate ducts by the pathologist (Dr. Bai) and the trained students and reviewed by D. Damaghi. We annotate ducts from WSIs of all three modalities. To ensure best characterization, we only identify ducts of >400 μms diameter, with visible myoepithelial layer and basement membrane. Following this, based on distance, each duct was annotated with four layers: adjacent stroma, oxidative/normoxia, hypoxic/hypoxia, and necrosis. Adjacent stroma was defined as the stroma up to 125 μms outside a given duct. Within the duct, necrosis was defined as any area containing dead cells, as identified by a lack of nuclei. Oxidative layer was defined as the area containing cells inside the duct within 125 μms of the basement membrane. Hypoxia was defined as the area containing cells inside the duct further than 125 μms from the basement membrane. The annotations were done for all 84 samples in the Bx cohort, and then were exported as standard GeoJSON files.

Co-registration. To characterize the interactions of different modalities from single-plexed slides, an alignment strategy was utilized. We register both CA9 and LAMP2b IHC slides towards the HE slides. A direct co-registering at the whole slide level with manual landmarks does not give us satisfactory alignment at each duct, due to the variable deformations across slides. We further co-register the slides in a duct-by-duct fashion. Using initially registered whole slides, and spatial proximity, we identify the corresponding ducts at the HE and 2 IHC slides. Next, we register both the CA9 duct and LAMP2b duct into the corresponding HE ducts. We use Virtual Alignment of pathology Image Series (VALIS), which provides a fully automated pipeline to register whole slide images (WSI) using rigid and/or non-rigid transformations (34). For each sample, we chose non-rigid registration and registered the ducts from CA9 and LAMP2b towards the reference HE ducts. The co-registration procedure and the qualitative results are shown in **Figure S4 and S5**. The co-registration provides a mapping of any cells detected in CA9 or LAMP2b towards a shared spatial domain, enabling the analysis of their interactions.

MODULE 2: Cell and niche detection

Cell detection. With the duct annotations in place, we automatically detect cells from the 2 IHCs and determine if they are positive in CA9 or LAMP2b based on their intensities. As we are only interested in intra-ductal cell niches, we only detect cells within each duct. For each IHC duct, we detect cells using Qupath watershed cell detection algorithm (25). Based on the intensity level, we categorize the cells into 4 groups: ‘Negative’, ’1+’, ’2+’, and ’3+’. The detection of cells within a duct is done by starDist (25,35) extension in Qupath on HE slide.

Graph construction for niche detection. After annotating all of the positive cells (i.e., CA9 or LAMP2b positive cells), they were mapped on HE slides, enabling us to detect niches on HE slides. Since there is a large amount of positive cells within each duct, with diverse spatial context and morphological features, we construct a graph with these cells by connecting cells whose distances are smaller than a certain threshold and detect connected components of the graph as representatives of cells living in “niches”. Multiple thresholds have been experimented and an optimum value is selected based on performance. Each positive cell niche is supposed to have a similar eco-evo phenotype and be spatially coherent. Therefore, we overlay both CA9 positive and LAMP2b positive cells into the same domain as an approximation of the local eco-evo cell distribution (**Figure S6**). This gives us the opportunity to measure their interaction via spatial statistical functions as defined later. Based on the same principle, we use cell morphological features extracted from HE within the region of each niche to characterize the niche.

MODULE 3: niche characterization and feature extraction

Once niches are detected. We extracted both spatial and morphological features to characterize them. To describe the spatial interaction patterns, we utilized various spatial functions as features. We also extract cell features consisting of morphology features and texture features that are commonly adopted in HE image analysis.

Cellular features. For cellular features we measured both morphological and texture features. The morphological features include area, eccentricities, circularity, elongation, extent, major axis length, minor axis length, solidity and curvature. The texture features include angular second moment (ASM) of co-occurrence matrix, contrast, correlation, entropy, homogeneity and intensity. All features were calculated following the implementations in the sc-MTOP(36) package.

Although we do not have exact cell-to-cell correspondence between the cells within a niche and cells detected in HE, we still can aggregate morphological and texture features within the proxy of the cells part of a niche to characterize the niche. For each niche, we identify the concave hull region enclosing its eco-evo positive cells within a duct on HE slide. Next, we aggregate cell features across all HE-detected cells within the corresponding region. For each cell feature dimension, we calculated its mean, standard deviation, maximum, minimum, kurtosis and skewness.

Spatial features. We extract various spatial statistical functions (37) to characterize residingcells and their interactions inside habitats to define niches. These functions are listed below:

G Function: The G function, denoted as G(r), is the cumulative distribution function of nearest-neighbor distance. The G function provides insights into the clustering or dispersion behavior of the point pattern.

$$G(r) = P\{d(u, X\backslash u) \leq r |\, u \in X\}, d(\bullet)\ is\ the\ minimum\ distance$$

F Function: The F function, known as the empty space function, is the cumulative distribution function of the empty-space distance. The F function is commonly used to assess the regularity or inhibition patterns in point patterns.

$$F(r) = P\{d(u, X) \leq r\}, d(\bullet)\ is\ the\ minimum\ distance$$

K Function: Ripley's K function, denoted as K(r), is a measure of second-order intensity or spatial interaction. It assesses whether points tend to be more clustered or dispersed within a certain distance r compared to a CSR process. It considers both the distance and intensity of points to capture the clustering behavior of the point pattern.

$$K(r) = \frac{|W|}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \mathbf{1}\{d_{ij} \leq r\} e_{ij}(r), e_{ij}(\bullet)\ is\ the\ edge\ correction\ weight$$

L Function: L function is a variance stabled version of K function.

$$L(r) = \frac{\sqrt{K(r)}}{r}$$

We calculated G, F, and L functions in both univariate and multivariate fashions. For each of the functions, the distances between source cell and the target cells are considered. Univariate spatial functions sample source cells and target cells from the same type of cells while multivariate counterparts' sample from different types of cells. Univariate G, F, and L are calculated for the single-marker cell subsets, and multivariate G_cross and L_cross for different subsets such as CA9-LAMP2b. 'Gest' function and 'Fest' function from 'spatstat' R package were used with Kaplan-Meier estimator(38), and 'Lest' function was used with isotropic correction(39,40).

MODULE 4: Diagnostic risk estimation with pattern proportion

In the last module, we train a classifier using these niches to predict whether a patient will be "upstaged" or "pure DCIS". This establishes the diagnostic power of these niches. A direct aggregation of niche information within each sample/patient is not sufficient. Tumor microenvironment is heterogeneous, and niches demonstrate diverse spatial and morphological behavior. To account for the diversity, we will focus on how different niches are distributed across a sample. We show that the distributions of different niches essentially characterize the tumor ecology in a much more refined manner compared with previous distance-based definitions of hypoxia/oxidative layers.

One technical challenge is that the niche features computed in the previous module are high dimensional and the niche features are diversely distributed. We propose to first find a simplified distributional description of the niches, and then use the simplified description for prediction. First, we cluster the niches into different sub-classes based on their features. The clustering is carried out using K-means clustering with a tunable parameter k. Once the niche sub-classes are determined. We use their distribution on each sample to predict its upstaged/pure DCIS status. The prediction power of the classifier sheds light on the diagnostic power of the niches and their spatial and cellular features. Five-fold cross-validation was employed, with one fold designated as the test set in each run. This approach prevents\s data leakage and helps mitigate overfitting.

To understand the contribution of each feature to the prediction model, we employed SHAP (SHapley Additive exPlanations) analysis. SHAP is a unified approach to interpreting machine learning models by assigning each feature an importance value for a particular prediction. In our study, SHAP values were computed for the features representing the proportions of different patterns within the niches. By calculating the SHAP values, we could determine the impact of each feature on the model's output, thereby identifying the most influential patterns that contribute to predicting DCIS upstaging. This step is crucial for ensuring the transparency and interpretability of the machine learning model.

Furthermore, we select features that are highly relevant to the sub-classes using different approaches including covariance, mutual information scoring and maximum relevance minimum redundancy (mRMR)(41) and choose the features identified by both approaches. **Figure 4C** shows the gradient map of each of these features on niches in the latent space.

Niche distribution for diagnosis. After assigning each duct to its sub-class, we aggregate across all niches of each sample and use its sub-class distribution to characterize this sample. Assuming k niche sub-classes, each sample has a k dimensional histogram to describe its niche sub-class distribution. We call this the niche distributional (Nbd-Dist) feature. We trained a classifier to predict whether a sample is pure DCIS or upstage. Repeating the iteration 10 times and comparing the mean area under curve (AUC) on the test set. The classifier types experimented include lightGBM, soft vector machine (SVM), logistic regression and random forest, and the random forest classifier yields the best performance.

Data Availability

The data generated in this study are available within the article and its supplementary data files. All the staining and annotations are also deposited in the physical sciences in oncology network (PSON).

Results:

Sample curation and cohort building

We built a retrospective cohort from 84 patients with histologically confirmed DCIS on core biopsy, followed by surgical excision, with available FFPE blocks at both Bx and Ex from Moffitt Cancer Center Biobank. The cohort has two arms: the first one is pure DCIS including the patient diagnosed with DCIS at both Bx and Ex. The second arm includes the upstaged group with DCIS at Bx and IDC at Ex (**Figure 1B**). HE stained slides of DCIS biopsy cores were retrieved from both the biobank core at the Moffitt Cancer Center tissue core and reviewed by a study pathologist (49). Then the selected blocks were pulled and sequentially cut for HE staining, CA9, and LAMP2b IHC staining. The HE and subsequent 2 IHC slides are digitally scanned using the Aperio XT® high-throughput slide scanner and housed on the web-based Aperio server/Spectrum database package. Upstage status was pulled from the electronic medical record and confirmed by our study pathologist from the Ex tissues (**Figure 1C**). All images were then segmented and annotated using Qupath supervised by study pathologist (25,49).

Annotation and eco-evolutionarily mapping of habitats at the individual duct level

We have shown previously that peri-luminal cells that are far (>0.125 - 0.160 mm) from a blood supply inhabit a microenvironment of hypoxia and acidosis (18,20,26). Thus, we created two simple annotation zones on HE slides based on oxygen diffusion distance representing oxygen defined habitats: i) hypoxic zone or habitat that is above 125 μms from the duct boundary, basement membrane, and ii) normoxic habitat that is the outer regions adjacent to the basement membrane (**Figure 2A**). We used the basement membrane as our zero point of reference. We also annotated necrotic zones inside the hypoxic habitats that also represent the anoxic habitat falling perfectly above 0.160 mm distance from basement membrane. Since adjacent stroma is also of interest to our group and others, we annotated adjacent stroma for each duct with binary scoring of 1 for having adjacent stroma or 0 for lacking it (**Supplementary Table 1**). To ensure a balanced representation of hypoxic and normoxic habitats, we excluded small ducts by establishing a duct size threshold of minimum 400 μms in diameter (or 200 μms radius) for manual annotation (**Figure S2**). After annotating all the ducts bigger than 200 μms of radius on HE slides, we expanded our annotations to other 2 consecutive IHC slides stained with CA9 and LAMP2b antibodies (**Figure 2B**). Subsequently, our pathologist, Dr. Bai, manually scored each duct for hypoxic and normoxic habitats based on CA9 and LAMP2b positivity using a scoring scale of 0–3 (**Supplementary Table 1**). Following this, positive cells in IHC slides were counted using Qupath (25), habitats were categorized into different classes based on the count of positive cells. The distribution of these habitat categories was compared between pure DCIS and upstaged groups (**Figure 2C, 2D, and S3)**. Using the Wilcoxon test, it was shown that there existed significant differences between pure DCIS and Upstaged group when habitats considered at the duct level. The tests were carried out for both hypoxic and oxidative layers for both CA9 (**Figure 2C and 2D)**, and for LAMP2b **(Figure S3**) as well as architecture, grade, lymphocytes, microcalcifications, and necrosis (**Supplementary Table 1**). As shown in Figure 2D, CA9 scoring within hypoxic habitats provides a much clearer distinction between pure DCIS and upstaged groups compared to the normoxic zone. Interestingly, CA9 did not show significant differences between the groups when analyzed at the whole duct or whole-slide level, as is traditionally done (**Figure S2B**). However, focusing on hypoxic or oxidative habitats revealed that CA9-positive cells are distributed differently between the two patient groups. This analysis underscores the value of examining fine-scale habitats within ducts. The improved performance of habitat-level scoring compared to whole-duct scoring highlights the necessity and significance of exploring the cellular composition and interactions within these microhabitats.

Mapping Metabolic Niches Within Habitats to Enhance Spatial Machine Learning Models

Previous analyses of hypoxic and normoxic habitats in breast cancer ducts were limited to scoring each biomarker individually, focusing solely on the count of positive cells within each habitat. To broaden the scope and incorporate interactions and relationships between these two eco-evolutionary marker-positive cells, a co-registration step was essential. This step enabled the creation of a virtual multiplex IHC (mIHC) by mapping cells onto a unified 2D reference space. HE slides were selected as the reference, and all IHC slides were registered onto this common framework. (**Figure S4**). Note that since our analysis is carried out duct-by-duct, it is not necessary to register the whole slide. Instead, for each duct, we register its IHC stainings to its HE staining. This ensures all the downstream analyses could be performed on the same HE slide coordinates system, providing consistency and precision in the spatial data integration. Then we used these mIHC images to define niches of cells that are positive for CA9, LAMP2b, or both. We hypothesized that niches within habitats characterized by both markers together would provide greater biological insight than analyzing each marker individually, given the established correlation between hypoxia and acid phenotypes. Then, we focus on the cell features such as nuclear morphology and texture and cell spatial features inside these niches to explore their predictive power on DCIS upstaging. As illustrated in **Figure 3**, we first map each positive cells to the reference HE slide using the co-registration described above. Then, by treating each positive cell as a node and connecting the cells within a distance threshold, we construct a cell-proximity graph out of mIHC positive cells whereby each connected component of this graph represents a continuous region or niche that is hypoxic, acidic, or both. The threshold is a tunable parameter that is optimized by the classifying power of the downstream analysis. And depending on the selection of the eco-evo markers, there can be CA9 positive niches, LAMP2b positive niches, or both CA9 and LAMP2b positive niches. We then developed a pattern differential analysis pipeline, which comprises two stages: First, the samples are clustered based on the features and classified into one of the clusters or patterns. Then for each patient, we calculate the proportion of each pattern, forming a distribution profile of the patterns.

By using these proportion features, we train a classifier aiming to predict the upstaging status. From this pipeline, we are able to predict the clinical outcome of a patient based on his/her spatially defined pattern distributions (**Figure 1C**). Then, to test the hypothesis that finer regions with biological meanings could provide better predictive power, we conduct a multi scale analysis performing a series of experiments using the same set of features and with the same pattern differential analysis pipeline at 3 different scales: duct, habitat, and niche (**Figure 1C**). At the habitat level, normoxic and hypoxic zones are analyzed independently. At the niche level, analyses are further refined to separately examine CA9-positive cells, LAMP2b-positive cells, and cells co-positive for both CA9 and LAMP2b.

For all the experiments, the biopsy dataset underwent 5-fold stratified cross-validation, where in each round, 4 folds served as the training dataset and 1-fold as the test dataset, with the goal of predicting the patients' clinical outcome at the time of biopsy. Upon comparing the mean accuracy score and the mean AUC score of all the classifiers, the niche level classifier yielded the best predictive results particularly under both metrics (**Table 1**). This result confirms that niche-based analysis outperforms our primary habitat analysis. The higher accuracy of the niche measurements may be implying the phenotype-based niche measurement is better than inferring habitat from oxygen diffusion rate measure based on the distance of the cells from basement membrane. Also, it is worth mentioning that oxygen habitat analysis is a rough estimate in our analysis since we do not know the exact location of the vasculature and their activity.

	Duct	**Habitat**		**Niche**		
		Normoxia	Hypoxia	CA9	LAMP2b	CA9 & LAMP2b
Accuracy	0.78 ± 0.06	0.86 ± 0.03	0.83 ± 0.06	0.82 ± 0.06	0.90 ± 0.03	0.90 ± 0.03
AUC	0.61 ± 0.08	0.67 ± 0.03	0.66 ± 0.10	0.64 ± 0.10	0.72 ± 0.07	0.74 ± 0.13

Table 1. **Performance scores of multi scale classifiers.** While habitat-level analysis enhanced performance, the niche-level classifier produced the most accurate predictive results.

Post analysis to reveal contributing features and prototype visualization on mIHC.

After identifying the best-performing classifier based on the AUC metric we employed SHAP (48) (Shapley Additive exPlanations) analysis to interpret the model by calculating SHAP values for each feature, specifically on the proportions of distinct patterns (**Figure 4b**). The pattern with the maximum SHAP value, identified as the most impactful, underwent further differential analysis to uncover features that significantly differentiated this pattern from others. This differential analysis employed methods including correlation analysis, mutual information (MI), and maximum relevance minimum redundancy (MRMR), which together identified Area_min, Perimeter_min, AreaBbox_min, and F_0 <= r < 10 as the top distinguishing features for Pattern 5 (**Figure 4c**). A prototype for Pattern 5, selected based on its alignment with the mean values of these features, was visualized to illustrate its characteristics (**Figure 4d**). Using a multi-scale analytical approach, we integrated spatial interactions of CA9-positive and LAMP2b-positive cells into the machine learning pipeline to distinguish between pure DCIS and progressed DCIS. Niche-level analysis yielded the highest accuracy and AUC, emphasizing the importance of fine-scale regions in predicting clinical outcomes. The use of SHAP analysis and differential analysis provided an interpretable framework to highlight influential patterns and features, such as Area_min and Perimeter_min, offering insights into the tumor microenvironment. This approach not only advanced our understanding of key spatial and morphological features but also demonstrated significant potential for precise diagnostic tools in clinical applications.

Discussion:

Ductal carcinoma in situ is the most prevalent type of precancer that can range from indolent to aggressive. DCIS lesions are highly heterogeneous in their intra- and inter- ductal physical microenvironments, genetics, and molecular expression patterns. They can be described as complete ecosystems containing habitats and niches including normal epithelial cells, pre-cancer cells, stromal cells, vasculature, structural proteins, signaling proteins and physical factors such as pH and oxygen concentration (18). These habitats and niches of micro-domains can contain unique mixtures of cells with physical and biochemical characteristics, with differential evolutionary potential and trajectories (27). The niches with similar mixtures of cells usually are also similar in their physiology and phenotypes mainly due to living in similar habitats. Our hypothesis is that knowledge of these niches and their habitats can potentially provide patient benefit by stratifying their tumor progress and therapeutic choices. However, tools and techniques are lacking to distinguish them. Proper tools and techniques can identify and define habitats and niches to map (pre-)cancer ecosystems to discriminate between the different types of DCIS to design the right treatment for breast cancer patients.

In this study, we argue that the overdiagnosis and overtreatment of DCIS stem from conventional frameworks that focus primarily on genetic signatures while neglecting the phenotypic heterogeneity within tumor ecosystems. Thus, we interpreted complex eco-evolutionary data of cancer cells within their niche using machine learning and pathomics, all framed within an innovative ecological and evolutionary dynamic model. Oxygen habitats are identified based on varying levels of perfusion and oxygenation, which are believed to play a crucial role in driving ecological diversity by changing cancer cells metabolism, creating new habitats, and enhancing tumor heterogeneity, ultimately leading to diverse evolutionary trajectories. (28, 29). Solid tumors often exhibit an impaired vascular system, leading to habitats within tumors that vary in hypoxia, nutrient deficiency, and acidity. These habitats can significantly influence the spatial selection of cellular phenotypes in distinct subregions. Inhabiting hypoxia, acidosis, and severe nutrient deprived habitats, face (pre-)cancer cells to strong selective pressures leading to divergence to novel phenotypes in population. These new phenotypes can reciprocally influence the microenvironment reshaping due to their new metabolic phenotypes resulting in a dynamically changing tumor ecosystem with multiple habitats. Therefore, the phenotype of the cells residing in these habitats can also be leveraged to define the habitats with a certain degree of accuracy. Previous research from our group and others demonstrated that cancer cells within breast ducts, exposed to chronic hypoxia and acidosis, develop adaptive mechanisms for survival in this challenging microenvironment including expression of CA9 or LAMP2b at the cell surface (18,20,30). However, none of these findings were used in a relevant translational study for biomarker discovery. In this study, we explore these biomarkers within an eco-evolutionary framework for the first time, using them as indicators of the metabolic state of cancer cells residing in a niche as part of oxygen habitats that may favor the selection of more aggressive phenotypes to predict the upstaging of DCIS. While a longitudinal study would indeed be a better study design for direct observation of evolutionary changes over time, our current cross-sectional approach enables us to capture a snapshot of the tumor microenvironment at two near time points, providing valuable insight into the conditions that distinguish DCIS from IDC. We recognize the assumption that the synchronous IDC microenvironment may contribute to the progression from DCIS to IDC. However, our study design allows us to test whether specific microenvironmental factors and related habitats and niche correlate with the presence of IDC, which can provide strong hypotheses for future longitudinal investigations. A future prospective or retrospective longitudinal (multiple long time points) study would indeed help distinguish whether these microenvironmental changes in tumor ecosystem locally belonged to habitats or niches can drive progression from DCIS to IDC or if IDC-induced those changes in the tumor ecosystem contribute to the synchronous DCIS phenotype.

In our curated retrospective cohort of 84 DCIS patients with histologically confirmed DCIS on core biopsy, we manually annotated 916 single ducts and more than 3000 habitats on all three slides and scored them at habitat levels. This unique detailed eco-evolutionary annotation can be used for future similar eco-evolutionary designed studies including stroma habitats. Our risk scoring system integrating principles of ecological-evolutionary dynamics with pathological imaging and molecular features of early-stage breast tumors showed improvement on prediction power of biomarkers alone and in combination.

We employed a 5-fold stratified cross-validation approach to ensure robust internal validation of our model. While this method helps mitigate overfitting and provides reliable performance estimates, we acknowledge the absence of an independent validation set, which is crucial for assessing the model's generalizability. The unique design of our cohort, which integrates specific ecological and microenvironmental factors, limits the availability of comparable external datasets for validation. As such, there is no current dataset with similar characteristics for cross-validation. We recognize this as a key limitation and emphasize that future studies should aim to validate the model on independent cohorts when such datasets become available. Furthermore, although our model achieved an AUC of 0.74, this performance is not yet sufficient for clinical translation. Additional efforts to refine the model and test it in larger, independent cohorts will be essential before its use in clinical practice can be considered. Interestingly, a recent approach using multiplex IF on DCIS cohort reached the same AUC(2). While both our study and the Risom et al. paper aim to leverage spatial relationships to predict DCIS progression, we would like to emphasize that the two approaches are fundamentally different in terms of the markers used. Risom et al. focused on a broad panel of markers, including those related to the stroma, immune cells, and tumor cells, which provide a comprehensive view of the tumor microenvironment. In contrast, our approach centers on eco-evolutionary markers derived from adaptation of cancer cells to physical microenvironment, specifically CA9 and LAMP2b, which are associated with hypoxia and tumor acidity and their spatial distribution, respectively. These differences reflect divergent hypotheses about the key drivers of DCIS progression. The fact that both studies report a similar AUC of 0.74, with the distinct marker sets and biological processes, suggests that our findings offer complementary insights into DCIS progression and combination of approaches might increase the accuracy.

Our study demonstrates the utility of eco-evolutionary principles in understanding DCIS progression. In our study, we proposed that specific tumor microenvironmental conditions, such as hypoxia and acidosis, are associated with phenotypic changes that may indicate DCIS progression. However, although we have shown previously that these microenvironments can cause aggressive phenotypes, we acknowledge that our findings here do not conclusively demonstrate that these environmental factors are causative agents in the transition from DCIS to IDC. Instead, our data suggest that these conditions could serve as biomarkers for identifying lesions that are more likely to be upstaged. However, the ability to define more refined cell phenotypes within each region of interest (ROI) could further enhance our analysis. If we can identify and characterize more detailed phenotypes, it would allow us to extract additional features that describe the spatial interactions of these phenotypes. This, in turn, could potentially improve the classifier's performance and make the results more interpretable. By capturing the intricate interactions between various cell types and their microenvironments, we could gain deeper insights into the ecological dynamics driving DCIS progression and improve predictive models for patient outcomes.

In recent years, there has been a growing trend towards adopting a "watchful waiting" approach for certain cases of DCIS, rather than immediate surgical excision(31,32). This strategy aims to reduce overtreatment by closely monitoring DCIS lesions that may not progress to invasive cancer. In this context, our upstaging predictions become particularly relevant. Identifying microenvironmental and phenotypic factors that indicate a higher likelihood of progression to IDC could help clinicians make more informed decisions about when to intervene and when to adopt a more conservative, observational approach. The ability to predict which DCIS cases are at higher risk of progressing to invasive disease would provide critical information for optimizing patient management, minimizing unnecessary treatments, and reducing the psychological and physical burdens associated with overtreatment(33). Further validation of these predictive models could therefore have important implications for guiding treatment strategies in the context of DCIS.

Lead contact

Further information and any related requests should be directed to and will be fulfilled by the lead contact Mehdi Damaghi (Mehdi.Damaghi@stonybrookmedicine.edu).

Acknowledgements

We gratefully acknowledge funding from Physical Sciences Oncology Network at the National Cancer Institute (grant U01CA261841), R01 grant R01CA272601, and R01CA249016. This work has been supported also in part by the Analytic Microscopy Core Facility at the H. Lee Moffitt Cancer Center & Research Institute, an NCI designated Comprehensive Cancer Center (P30-CA076292). We thank Dr. Alexander Borowsky, for reading our manuscript and constructive comments on pathological aspect as well as Dr. Liliana Davolos and Dr. David Ray for their guidelines on ecological and evolutionary dynamics principles used in the manuscript.

Author contributions

M.D. conceptualized and designed the research; Y.X., M.A., J.D.B., A.C., Y.C., M.D., performed the experiment and analysis; J.D.B. reviewed all the slides and scored them as the project pathologist; P.P., C.C., and M.D. contributed to results interpretation; and M.D. wrote the paper. All authors revised the paper.

Figure Captions

Figure 1. Ecological and evolutionary designed biomarkers of DCIS upstaging. A) Model of microenvironment-driven evolution of breast cancer from normal breast tissue to DCIS and IDC: Our schematic is overlaid on HE staining of breast cancer specimens at different stages of DCIS and IDC. Different patients may experience various types of evolutionary trajectory following different evolutionary models, including linear and branched progression from DCIS to IDC shown here. Note that these events are not sequential or stepwise. **B)** The patient cohort was curated from retrospective DCIS samples, with two sample collections at biopsy and excision. The main criterion was the diagnosis of DCIS at the biopsy stage. **C)** Eco-evolutionary designed- machine learning assisted pipeline to define cancer cell niches inside oxygen habitats in DCIS. ***i*)** Data preprocessing steps including duct annotation, cell detection and classification for HE and IHC slides, followed by co-registration to map IHC-identified cells onto the HE slides. ***ii*)**The analysis is carried out at multiple scales, namely duct, habitat and niche, from the largest to smallest. At each scale the nucleus morphology texture feature and spatial features are extracted. ***iii*)** The pattern differential analysis approach where the patterns are firstly identified and then the proportions of such patterns are used as features to predict the upstaging status of a patient.

Figure 2. **Eco-evolutionarily designed biomarker discovery to predict upstaging in DCIS. A)** Illustration of normoxic, hypoxic and necrotic habitats in a duct. **B)** Illustration of annotation and scoring on 2 IHCs and how cells are scored in each habitat. **C)** and **D)** Dot plots of counts of CA9 expression in each habitat per duct. Cells are scored 0 for 'negative' or '1+','2+','3+' for positive cells based on their intensity. Scoring was performed and analyzed separately for normoxic (oxidative) habitat (C) or hypoxic habitat (D). In the dot plot, each dot is a single duct. The color of dots reflects their score as follows: Blue = 0, yellow ='1+', orange ='2+', and red = '3+'. The number of dots reflects how many ducts were detected in each patient's biopsy with size bigger than 400 µms in diameter. The distribution in hypoxic habitat is significantly different between pure DCIS and upstaged groups in hypoxic habitats and not in oxygenated habitat. Data was analyzed using the Wilcoxon signed-rank test. The same graph is created for LAMP2b (supplementary fig. 2).

Figure 3. Niches are defined inside habitats from the hypoxia and acidosis markers expression. A) One sample duct from CA9 slide. Top: The original IHC slide. Middle: Cell detection and intensity-based classification using Qupath overlaid on the slide. Bottom: the graph constructed from the CA9 positive cells and the connected components of the graph (Niches) highlighted in different colors. **B)** The HE staining of the same duct as A. Top: The original HE slide. Middle: Duct annotation overlaid on the HE slide. Bottom: Co-registered CA9-positive niches mapped and overlaid on HE slides as mIHC to be able to extract HE features from CA9 positive niches. Note the orientation of HE and CA9 slide was opposite, and our co-registration technique successfully created a mIHC of the ducts with similar coordinates. The same approach was used for LAMP2b and the combination.

Figure 4. Post Analysis reveals the top contributing patterns and features. A) UMAP of the features of the niches, different colors represent different clusters(patterns) **B)** Top: The impact of each pattern on the classifying result, blue and red colors represent impact on pure DCIS and progressed predictions respectively, the proportion of pattern 5 has the greatest impact for both categories. Bottom: Using correlation, MI, and MRMR to obtain the most contributing features in the pattern 5 clustering phase, identifying a common feature set that includes 4 features: Area_min, Perimeter_min, AreaBbox_min, and F_0<=r<10. **C)** UMAP showing the value of the 4 identified features for different samples, and it can be seen that samples in the pattern 5 tend to have higher values in Area_min, Perimeter_min, AreaBbox_min and low values for F_0<=r<10. **D)** A niche belonging to pattern 5, it contains no small size cells and exhibits a relatively dispersed distribution.

References:

1. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481:306–13.

2. Risom T, Glass DR, Averbukh I, Liu CC, Baranski A, Kagel A, et al. Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. Cell. 2022;185:299–310.e18.

3. Maley CC, Aktipis A, Graham TA, Sottoriva A, Boddy AM, Janiszewska M, et al. Classifying the evolutionary and ecological features of neoplasms. Nat Rev Cancer. 2017;17:605–19.

4. Boutry J, Tissot S, Ujvari B, Capp J-P, Giraudeau M, Nedelcu AM, et al. The evolution and ecology of benign tumors. Biochim Biophys Acta Rev Cancer. 2022;1877:188643.

5. Amend SR, Pienta K. Abstract 2884: Tumor-driven eutrophication of the tumor ecosystem selects for cancer cell clones that overcome evolutionary inertia leading to increased metastatic capacity. Cancer Res. American Association for Cancer Research; 2015;75:2884–2884.

6. Damaghi M, Mori H, Byrne S, Xu L, Chen T, Johnson J, et al. Collagen production and niche engineering: A novel strategy for cancer cells to survive acidosis in DCIS and evolve. Evol Appl. 2020;13:2689–703.

7. Lipinski KA, Barber LJ, Davies MN, Ashenden M, Sottoriva A, Gerlinger M. Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. Trends Cancer Res. 2016;2:49–63.

8. Giaquinto AN, Sung H, Miller KD, Kramer JL, Newman LA, Minihan A, et al. Breast Cancer Statistics, 2022. CA Cancer J Clin. 2022;72:524–41.

9. Strand SH, Rivero-Gutiérrez B, Houlahan KE, Seoane JA, King LM, Risom T, et al. Molecular classification and biomarkers of clinical outcome in breast ductal carcinoma in situ: Analysis of TBCRC 038 and RAHBT cohorts. Cancer Cell. 2022;40:1521–36.e7.

10. Lehman CD, Arao RF, Sprague BL, Lee JM, Buist DSM, Kerlikowske K, et al. National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. Radiology. 2017;283:49–58.

11. Lips EH, Kumar T, Megalios A, Visser LL, Sheinman M, Fortunato A, et al. Genomic analysis defines clonal relationships of ductal carcinoma in situ and recurrent invasive breast cancer. Nat Genet. 2022;54:850–60.

12. Sarhadi S, Salehzadeh-Yazdi A, Damaghi M, Zarghami N, Wolkenhauer O, Hosseini H. Omics Integration Analyses Reveal the Early Evolution of Malignancy in Breast Cancer. Cancers [Internet]. 2020;12. Available from: http://dx.doi.org/10.3390/cancers12061460

13. Heselmeyer-Haddad K, Berroa Garcia LY, Bradley A, Ortiz-Melendez C, Lee W-J, Christensen R, et al. Single-cell genetic analysis of ductal carcinoma in situ and invasive breast cancer reveals enormous tumor heterogeneity yet conserved genomic imbalances and gain of MYC during progression. Am J Pathol. 2012;181:1807–22.

14. Hanna WM, Parra-Herran C, Lu F-I, Slodkowska E, Rakovitch E, Nofech-Mozes S. Ductal carcinoma in situ of the breast: an update for the pathologist in the era of individualized risk assessment and tailored therapies. Mod Pathol. 2019;32:896–915.

15. Carmeliet P, Jain RK. Principles and mechanisms of vessel normalization for cancer and other angiogenic diseases. Nat Rev Drug Discov. 2011;10:417–27.

16. Wu B, Liu D-A, Guan L, Myint PK, Chin L, Dang H, et al. Stiff matrix induces exosome secretion to promote tumour growth. Nat Cell Biol. 2023;25:415–24.

17. Persi E, Duran-Frigola M, Damaghi M, Roush WR, Aloy P, Cleveland JL, et al. Systems

analysis of intracellular pH vulnerabilities for cancer therapy. Nat Commun. 2018;9:2997.

18. Damaghi M, West J, Robertson-Tessi M, Xu L, Ferrall-Fairbanks MC, Stewart PA, et al. The harsh microenvironment in early breast cancer selects for a Warburg phenotype. Proc Natl Acad Sci U S A [Internet]. 2021;118. Available from: http://dx.doi.org/10.1073/pnas.2011342118

19. Lobo RC, Hubbard NE, Damonte P, Mori H, Pénzváltó Z, Pham C, et al. Glucose Uptake and Intracellular pH in a Mouse Model of Ductal Carcinoma In situ (DCIS) Suggests Metabolic Heterogeneity. Front Cell Dev Biol. 2016;4:93.

20. Damaghi M, Tafreshi NK, Lloyd MC, Sprung R, Estrella V, Wojtkowiak JW, et al. Chronic acidosis in the tumour microenvironment selects for overexpression of LAMP2 in the plasma membrane. Nat Commun. 2015;6:8752.

21. Ordway B, Swietach P, Gillies RJ, Damaghi M. Causes and Consequences of Variable Tumor Cell Metabolism on Heritable Modifications and Tumor Evolution. Front Oncol. 2020;10:373.

22. Gillies RJ, Verduzco D, Gatenby RA. Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. Nat Rev Cancer. 2012;12:487–93.

23. Ibrahim-Hashim A, Robertson-Tessi M, Enriquez-Navas PM, Damaghi M, Balagurunathan Y, Wojtkowiak JW, et al. Defining Cancer Subpopulations by Adaptive Strategies Rather Than Molecular Properties Provides Novel Insights into Intratumoral Evolution. Cancer Res. 2017;77:2242–54.

24. Damaghi M, Gillies R. Phenotypic changes of acid-adapted cancer cells push them toward aggressiveness in their evolution in the tumor microenvironment. Cell Cycle. 2017;16:1739–43.

25. 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.

26. Freischel AR, Damaghi M, Cunningham JJ, Ibrahim-Hashim A, Gillies RJ, Gatenby RA, et al. Frequency-dependent interactions determine outcome of competition between two breast cancer cell lines. Sci Rep. 2021;11:4908.

27. Jardim-Perassi BV, Huang S, Dominguez-Viqueira W, Poleszczuk J, Budzevich MM, Abdalah MA, et al. Multiparametric MRI and Coregistered Histology Identify Tumor Habitats in Breast Cancer Mouse Models. Cancer Res. 2019;79:3952–64.

28. Sobhani F, Muralidhar S, Hamidinekoo A, Hall AH, King LM, Marks JR, et al. Spatial interplay of tissue hypoxia and T-cell regulation in ductal carcinoma in situ. NPJ Breast Cancer. 2022;8:105.

29. Compton ZT, Mallo D, Maley CC. Stronger together: Cancer clones cooperate to alleviate growth barriers in critical cancer progression transitions. Cancer Res. American Association for Cancer Research (AACR); 2023;83:4013–4.

30. Chafe SC, McDonald PC, Saberi S, Nemirovsky O, Venkateswaran G, Burugu S, et al. Targeting hypoxia-induced carbonic anhydrase IX enhances immune-checkpoint blockade locally and systemically. Cancer Immunol Res. American Association for Cancer Research; 2019;7:1064–78.

31. Ryser MD, Worni M, Turner EL, Marks JR, Durrett R, Hwang ES. Outcomes of active surveillance for ductal carcinoma in situ: A computational risk analysis. J Natl Cancer Inst. Oxford University Press (OUP); 2016;108:djv372.

32. Glencer AC, Miller PN, Greenwood H, Maldonado Rodas CK, Freimanis R, Basu A, et al. Identifying good candidates for active surveillance of ductal carcinoma in situ: Insights from a large neoadjuvant endocrine therapy cohort. Cancer Res Commun. American Association for Cancer Research (AACR); 2022;2:1579–89.

33. Fortunato A, Mallo D, Cisneros L, King LM, Khan A, Curtis C, et al. Evolutionary Measures Show that Recurrence of DCIS is Distinct from Progression to Breast Cancer. medRxiv [Internet]. 2024; Available from: http://dx.doi.org/10.1101/2024.08.15.24311949

34. Gatenbee CD, Baker A-M, Prabhakaran S, Swinyard O, Slebos RJC, Mandal G, et al. Virtual alignment of pathology image series for multi-gigapixel whole slide images. Nat Commun. 2023;14:4502.

35. Schmidt U, Weigert M, Broaddus C, Myers G. Cell Detection with Star-Convex Polygons. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing; 2018. page 265–73.

36. Zhao S, Chen D-P, Fu T, Yang J-C, Ma D, Zhu X-Z, et al. Single-cell morphological and topological atlas reveals the ecosystem diversity of human breast cancer. Nat Commun. 2023;14:6796.

37. Baddeley A, Rubak E, Turner R. Spatial Point Patterns: Methodology and Applications with R. CRC Press; 2015.

38. Baddeley A, Gill R. Kaplan-Meier estimators of interpoint distance distributions for spatial point processes. IEEE Trans Inf Theory [Internet]. 1993; Available from: https://ir.cwi.nl/pub/5247/05247D.pdf

39. Ohser J. On estimators for the reduced second moment measure of point processes. Series Statistics. Taylor & Francis; 1983;14:63–71.

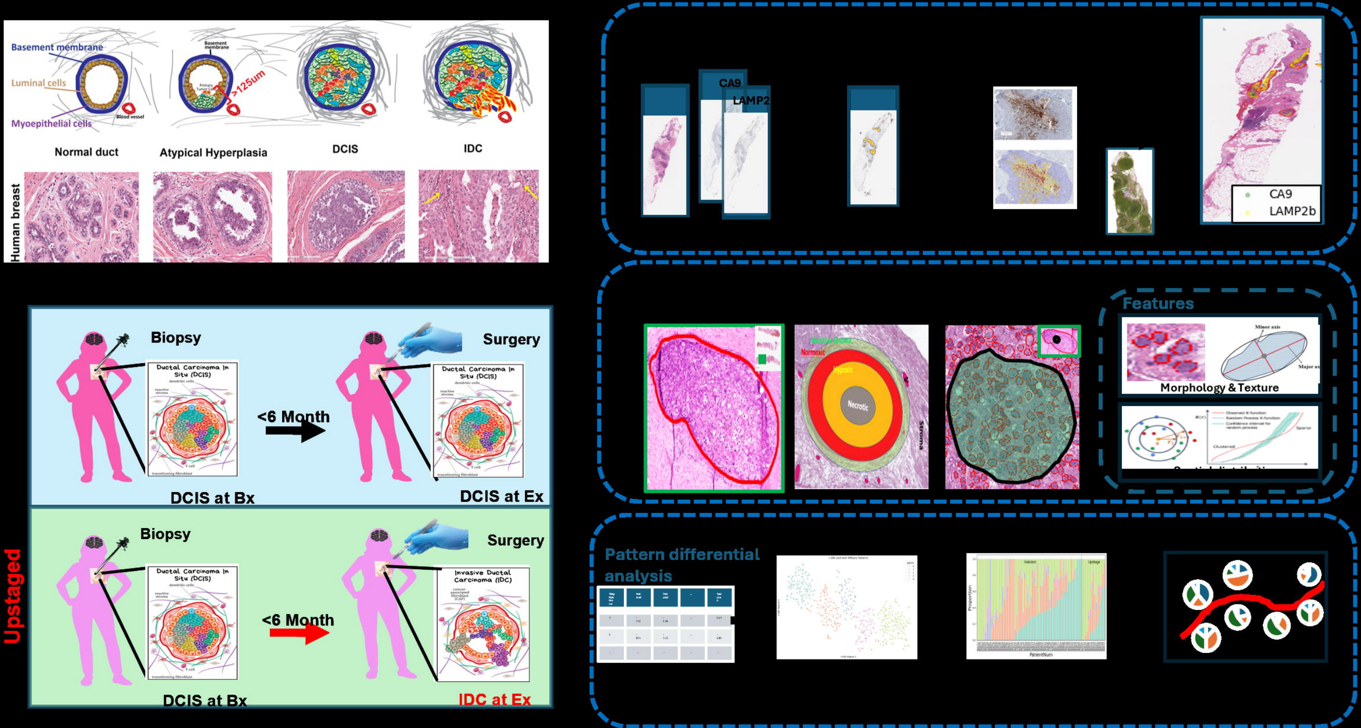
40. Kendall WS. Stochastic Geometry: Likelihood and Computation. Routledge; 2019.

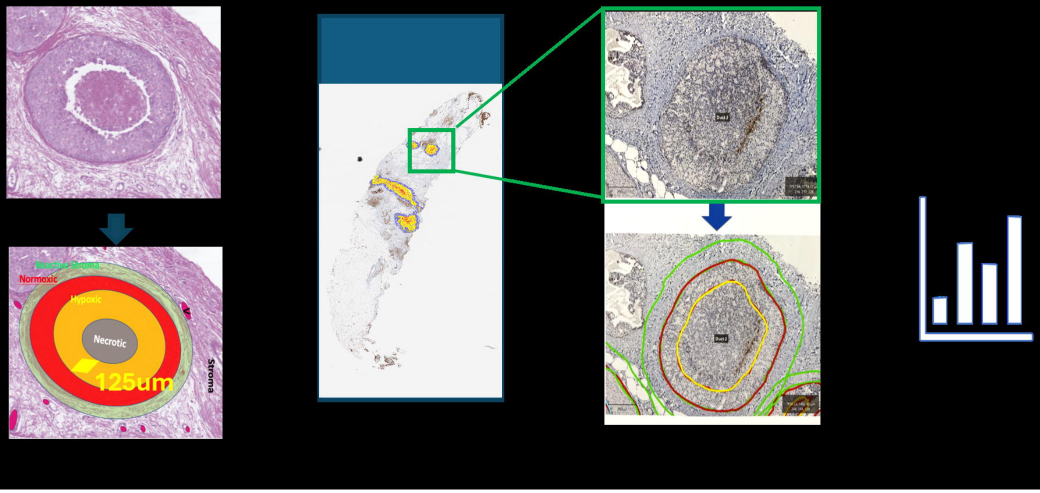
41. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27:1226–38.

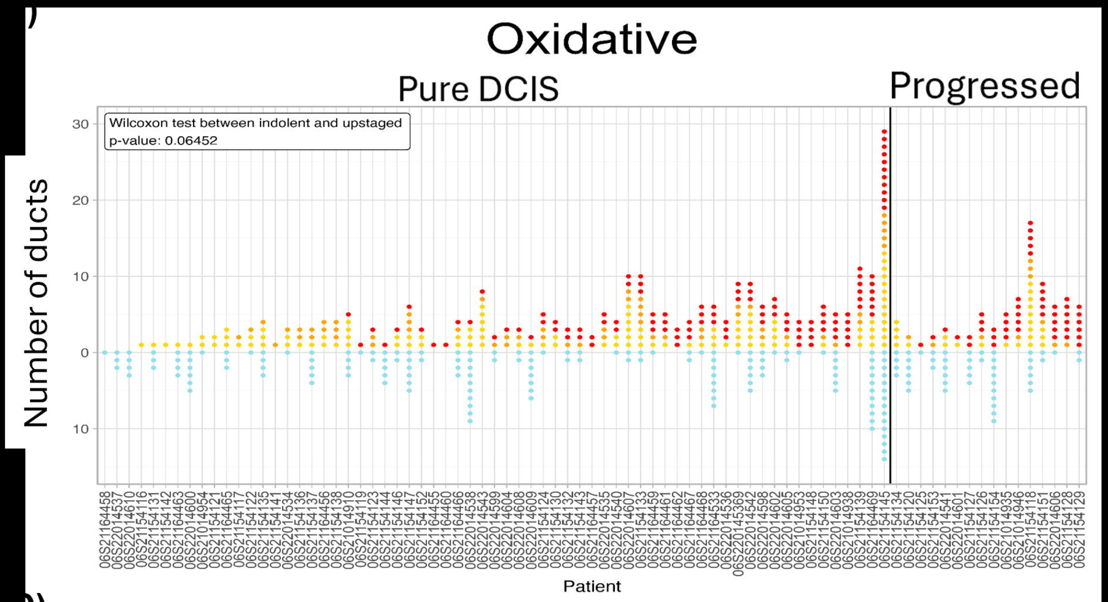
48. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Proc. Adv. Neural Inf. Process. Syst.* 30 (2017).

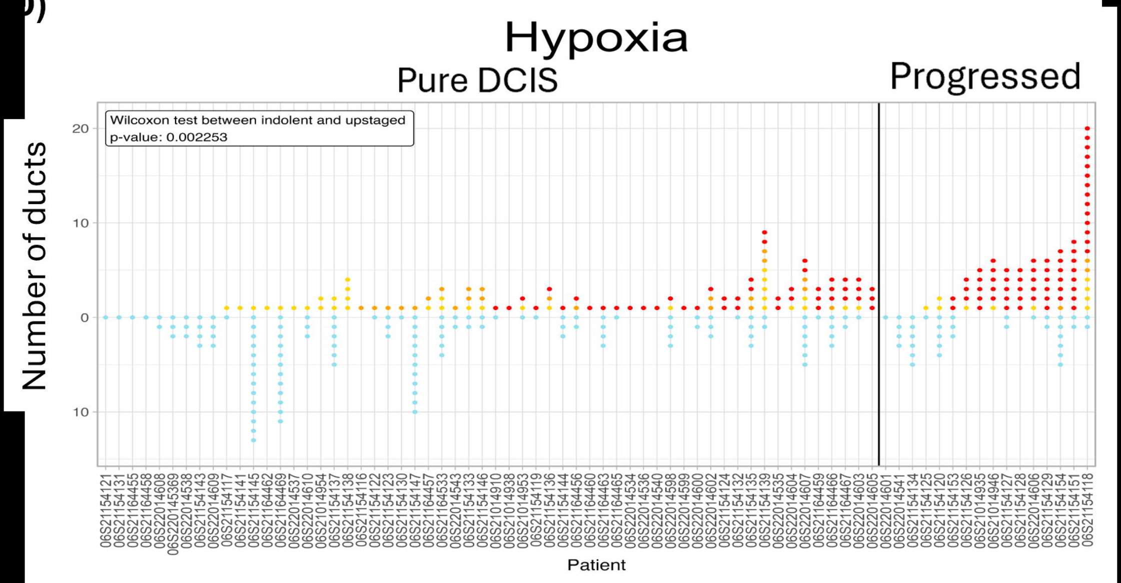
49. Mayfield, J. D., Ataya, D., Abdalah, M., Stringfield, O., Bui, M. M., Raghunand, N., Niell,

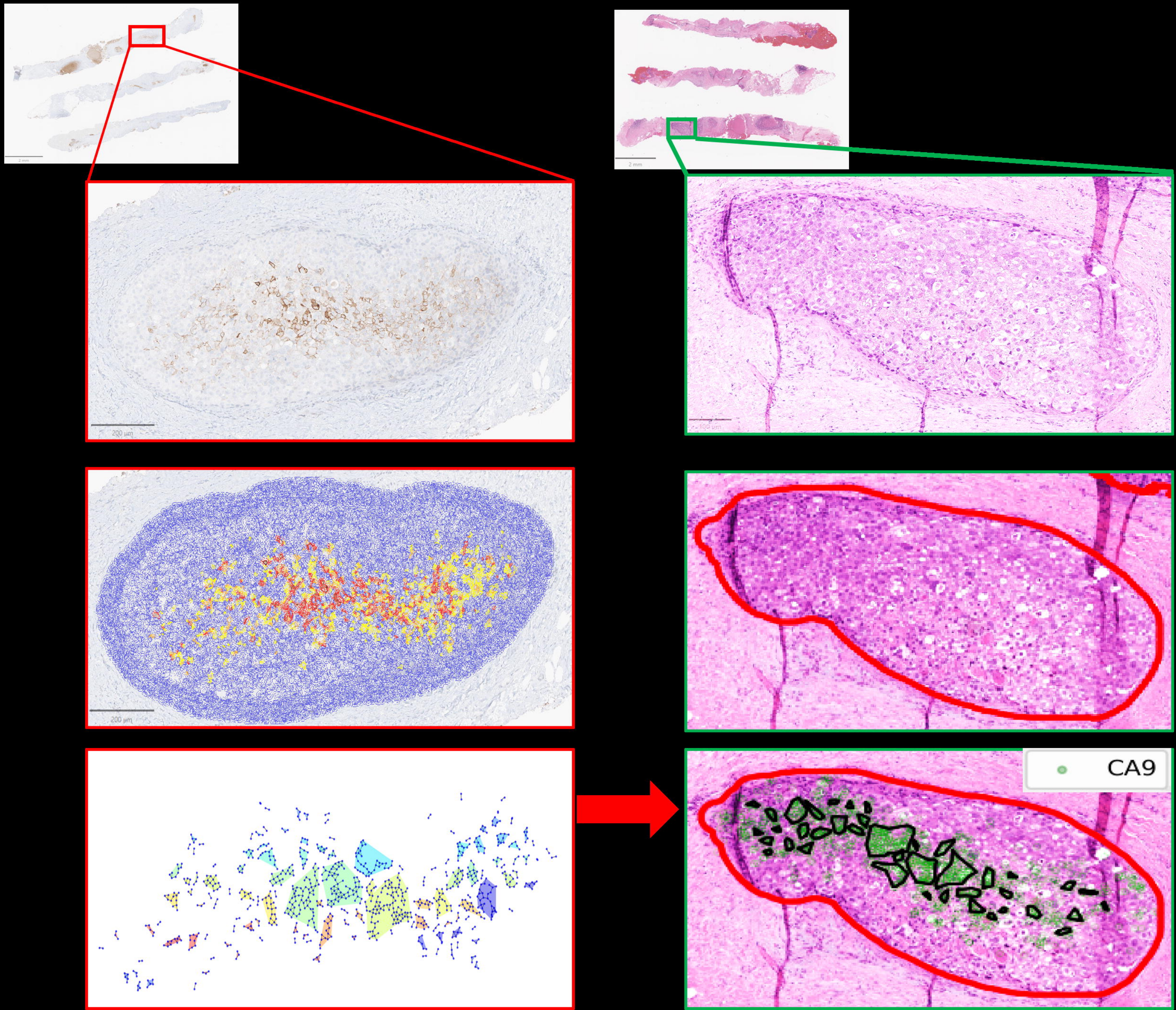
B., & El Naqa, I. Presurgical upgrade prediction of DCIS to invasive ductal carcinoma using time-dependent deep learning models with DCE MRI. Radiology: Artificial Intelligence 6, e230057 (2024).

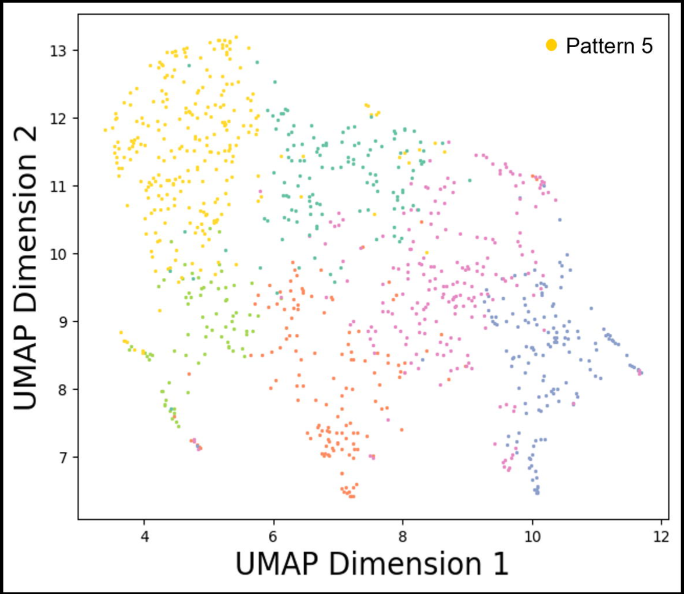

Basement membrane
Luminal cells
Myoepithelial cells
Blood vessel
Basement membrane
Primary Tumor Cell
>125um
Normal duct
Atypical Hyperplasia
DCIS
IDC
Human breast
Biopsy
Surgery
Ductal Carcinoma In Situ (DCIS)
<6 Month
DCIS at Bx
DCIS at Ex
Upstaged
Invasive Ductal Carcinoma (IDC)
IDC at Ex
CA9
LAMP2
CA9
LAMP2b
Normoxic
Hypoxic
Necrotic
Stroma
Features
Morphology & Texture
Minor axis
Major ax
Clustered
Sparse
Observed K-function
Random Process K-function
Confidence interval for random process
Pattern differential analysis
Proportion
PatientNum
Indolent
Upstage

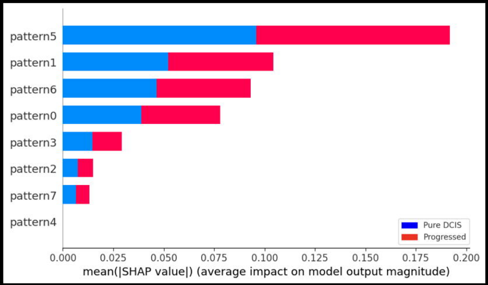

Reactive Stroma
Normoxic
Hypoxic
Necrotic
125um
Stroma
V
Duct 2
Duct 2

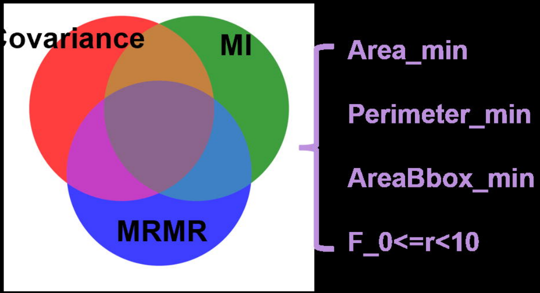

)
Oxidative
Pure DCIS
Progressed
Wilcoxon test between indolent and upstaged
p-value: 0.06452
Number of ducts
30
20
10
0
10
06S21164458
06S22014537
06S22014610
06S21154116
06S21154131
06S21154142
06S21164463
06S22014600
06S21014954
06S21154121
06S21164465
06S21154117
06S21154122
06S21154135
06S21154141
06S22014534
06S21154136
06S21154137
06S21164456
06S21154138
06S21014910
06S21154119
06S21154123
06S21154144
06S21154146
06S21154147
06S21154152
06S21164455
06S21164460
06S21164466
06S22014538
06S22014543
06S22014599
06S22014604
06S22014608
06S22014609
06S21154124
06S21154130
06S21154132
06S21154143
06S21164457
06S22014535
06S22014540
06S22014607
06S21154133
06S21164459
06S21164461
06S21164462
06S21164467
06S21164468
06S21164533
06S22014536
06S220145369
06S22014542
06S22014598
06S22014602
06S22014605
06S21014953
06S21154148
06S21154150
06S22014603
06S21014938
06S21154139
06S21164469
06S21154145
06S21154134
06S21154120
06S21154125
06S21154153
06S22014541
06S22014601
06S21154127
06S21154126
06S21154154
06S21014935
06S21014946
06S21154118
06S21154151
06S22014606
06S21154128
06S21154129
Patient

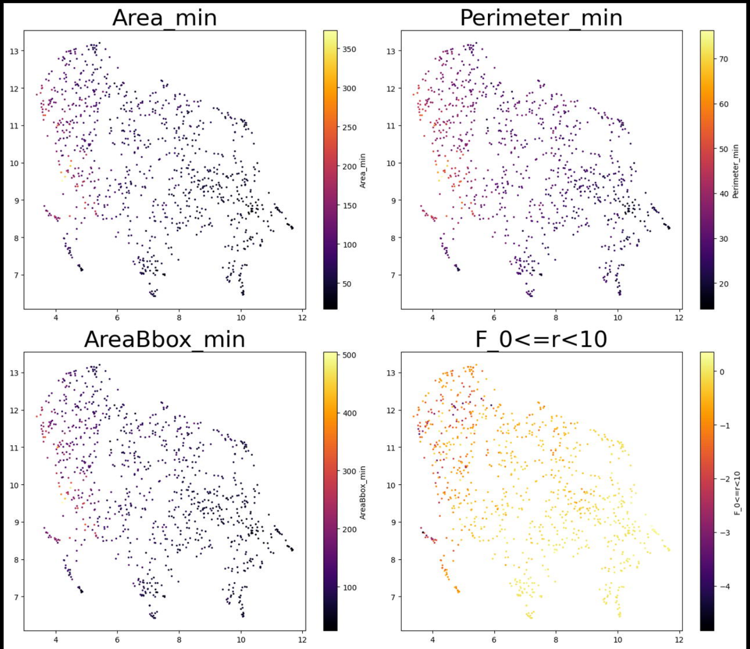

)
Hypoxia
Pure DCIS
Progressed
Wilcoxon test between indolent and upstaged
p-value: 0.002253
Number of ducts
20
10
0
10
06S21154121
06S21154131
06S21164455
06S21164458
06S22014608
06S220145369
06S22014538
06S21154143
06S22014609
06S21154117
06S21154141
06S21154145
06S21164462
06S21164469
06S22014537
06S22014610
06S21014954
06S21154137
06S21154138
06S21154116
06S21154122
06S21154123
06S21154130
06S21154147
06S21164457
06S22014533
06S22014543
06S21154133
06S21154146
06S21014910
06S21014938
06S21014953
06S21154119
06S21154136
06S21154144
06S21164456
06S21164460
06S21164463
06S21164465
06S22014534
06S22014536
06S22014540
06S22014598
06S22014599
06S22014600
06S22014602
06S21154124
06S21154132
06S21154135
06S21154139
06S22014535
06S22014604
06S22014607
06S21164459
06S21164466
06S21164467
06S22014603
06S22014605
06S22014601
06S22014541
06S21154134
06S21154125
06S21154120
06S21154153
06S21154126
06S21014935
06S21014946
06S21154127
06S21154128
06S22014606
06S21154129
06S21154154
06S21154151
06S21154118
Patient


2 mm
200 µm
200 µm
2 mm
100 µm
CA9


Pattern 5
UMAP Dimension 1
UMAP Dimension 2
4
6
8
10
12
7
8
9
10
11
12
13


pattern5
pattern1
pattern6
pattern0
pattern3
pattern2
pattern7
pattern4
Pure DCIS
Progressed
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
mean(|SHAP value|) (average impact on model output magnitude)


Covariance
MI
MRMR
Area_min
Perimeter_min
AreaBbox_min
F_0<=r<10


Area_min
Perimeter_min
AreaBbox_min
F_0<=r<10