Eco-evolutionary Guided Pathomic 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

[*] Corresponding authors: Chao Chen: chao.chen.1@stonybrook.edu, Mehdi Damaghi: mehdi.damaghi@stonybrookmedicine.edu

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. Although there has been extensive research on genetic and epigenetic causes of breast carcinogenesis, none of these studies have successfully identified a biomarker for the progression and/or upstaging of DCIS. 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 in our DCIS cohort of 84 patients. Then, we identify cancer cells with metabolic phenotypes attributed to their habitat conditions, such as the expression of CA9 indicating hypoxia responding phenotype, and LAMP2b indicating a hypoxia-induced 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 indolent and upstaged DCIS increased significantly. Second, using eco-evolutionary guided computational and digital pathology techniques, we discovered distinct spatial patterns of these biomarkers and used the distribution of such patterns to predict patient upstaging. The patterns were characterized by both cellular features 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 eco-evolution dynamics in predicting cancer progression.

Keywords:

Breast cancer, tumor ecology and evolution, DCIS, Eco-evolutionary biomarkers, Metabolic phenotypes, Habitats, Pathomic, Machine learning, digital pathology

Introduction:

In recent years, the understanding that cancer is a dynamic ecological and evolutionary process has become deeply entrenched [1,23]. 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 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 within tissues [5,6]. This ecological perspective of cancer is crucial for discovering the natural processes 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 formalin-fixed paraffin-embedded (FFPE) 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 upstage 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 upstaged to IDC[9].

DCIS is a heterogeneous group of neoplastic lesions confined to the mammary ducts. The confinement of proliferating cancer cells inside the duct and growth of 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 upstaging 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 clonogensis[11]. These new findings emphasize the extraordinary heterogeneity in genotype and phenotypic plasticity in breast cancer that must be studied in 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 "cell habitats" and their interactions in the tumor ecosystem. Tumor evolution requires phenotypic diversity within a population undergoing microenvironmental selection forces [21]. Cells that adapt in response to 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], 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

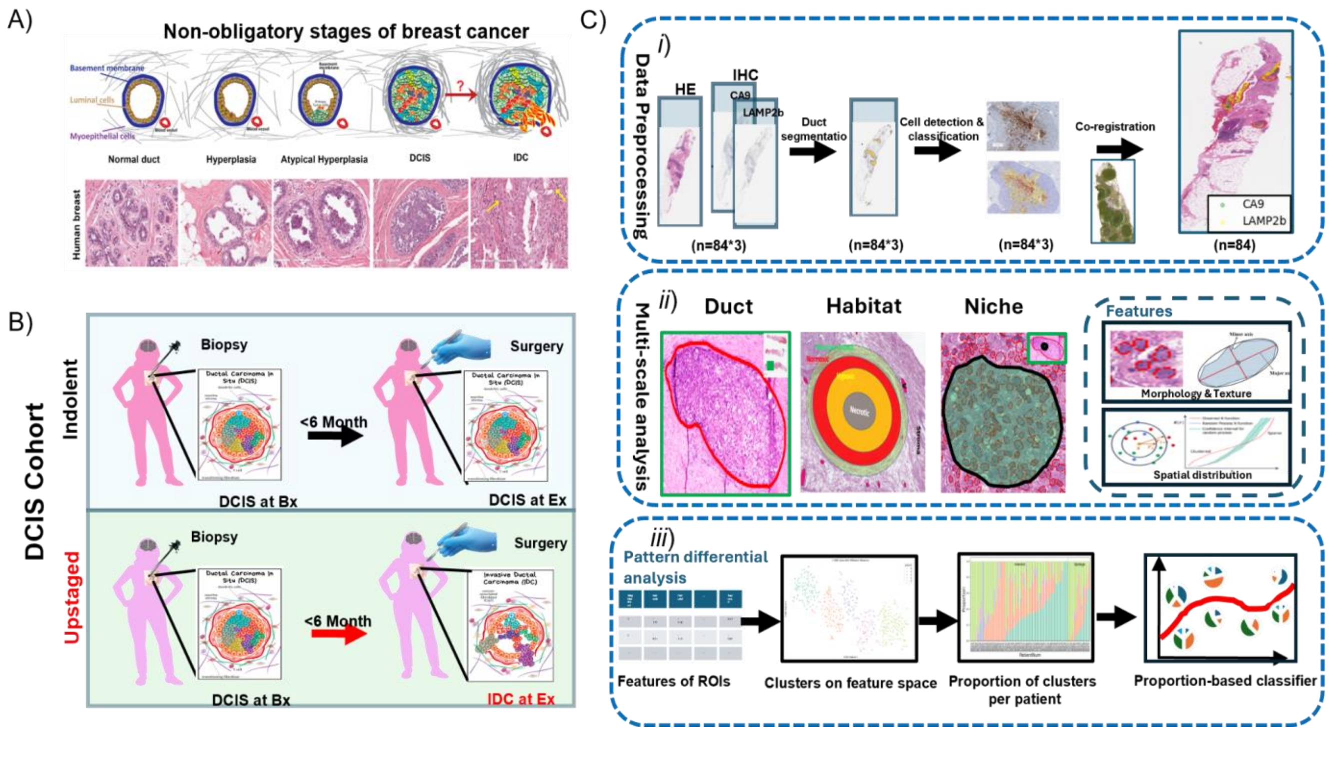

Figure 1. Ecological and evolutionary designed biomarkers of DCIS upstaging. A) Model of microenvironment-driven evolution of breast cancer from normal breast tissue to DCIS and IDC: Our schematic is overlaid on HE staining of breast cancer specimens at different stages of DCIS and IDC. Different patients may experience various types of evolutionary trajectory following different evolutionary models, including linear and branched progression from DCIS to IDC shown here. **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 coregistration to map IHC-identified cells onto the HE slide. ***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.

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[6,17,20,23,24,18]. 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. This 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 HE, CA9 and LAMP2b. We manually annotated ducts bigger than 400 μms in diameter. The 200 um 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 hypoxic habitats. Multiple spatial functions and spatial entropies were used to describe the spatial patterns of the cell groups (niche and micro-niche). 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 micro-niches with spatial and pathomic 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 immunohistological images combined with the tumor's eco-evolution dynamics and underlying molecular pathophysiology can significantly improve predicting the evolutionary trajectory of that 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 indolent 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

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. The cohort has two arms: the first one is indolent 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**). Hematoxylin and eosin (HE) stained slides of DCIS biopsy cores were retrieved from both the biobank core at Stony Brook University and the Moffitt Cancer Center tissue core and reviewed by our study pathologist. Then the selected blocks were pulled and sequentially cut for HE staining and 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 approved by our study pathologist from the Ex tissues (**Figure 1C**). All images were then segmented and annotated using Qupath supervised by study pathologist [25].

Annotation of eco-evolutionarily defined 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 lactic acidosis [18,20,26]. Thus we created two simple annotation zones on HE slides based on O2 diffusion distance representing oxygen 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 reactive stroma is also of interest to our group and others, we annotated reactive stroma for each duct with binary scoring of 1 for having reactive stroma or 0 for lacking it (**Supplementary Table 1**). To ensure a balanced representation of hypoxic and normoxic habitats, we established a duct size threshold of minimum 400 µms in diameter (or 200 µms radius) for manual annotation (**Figure S1**). After annotating all the ducts bigger than 200 ums of radius on HE slides, we expanded our annotations to other 2 consecutive IHC slides stained with CA9 and LAMP2b antibodies (**Figure 2B**). Then our pathologist manually scored each duct for both hypoxic and normoxic habitats separately for positivity of CA9 (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, and the count of each category of habitats for each patient was then compared between the upstaged and indolent patient groups.

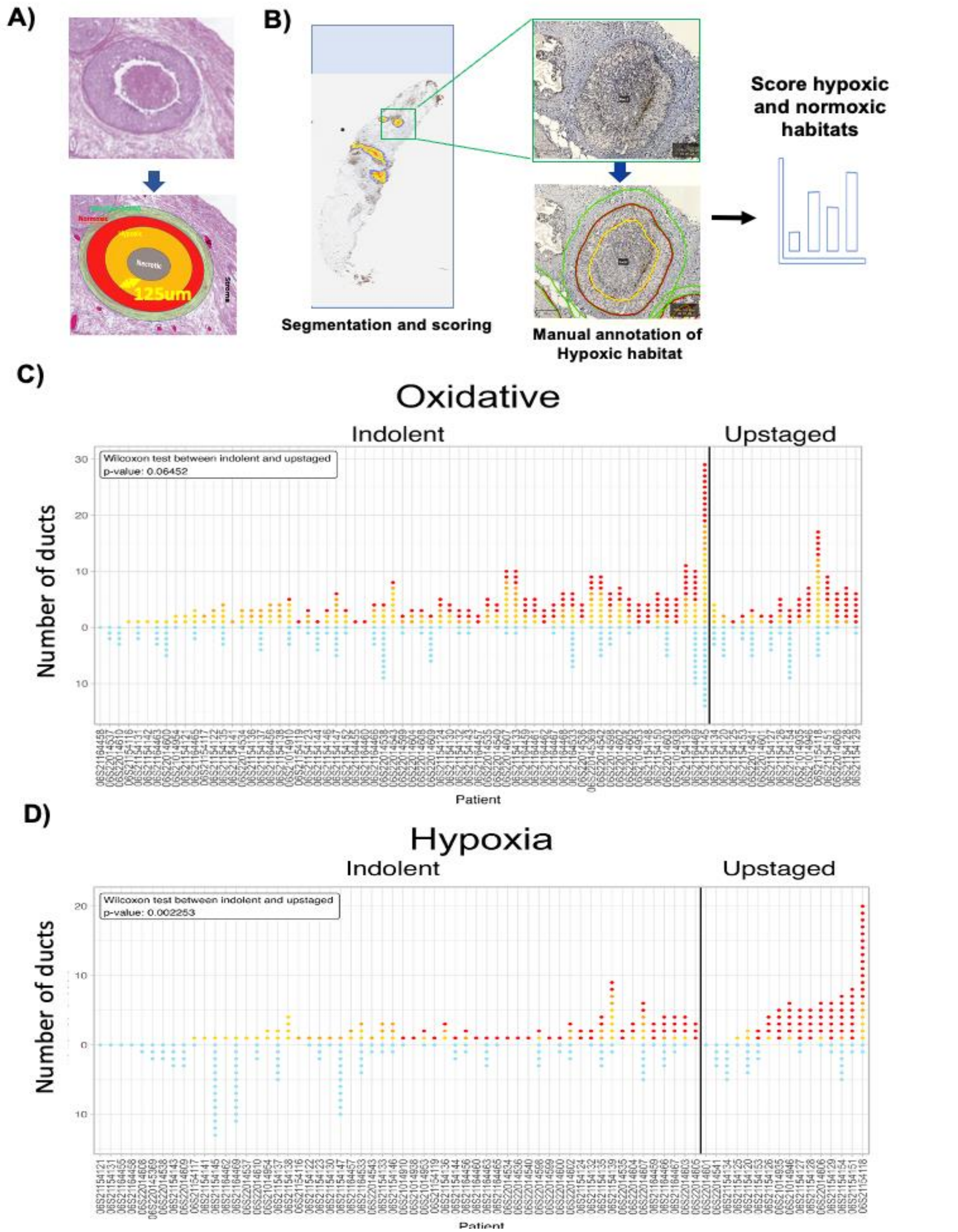

Figure 2. **Eco-evolutionarily designed biomarker discovery to predict upstaging in DCIS. A)**

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 habitats. **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 reflect 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 um in diameter. The distribution in hypoxic habitat is significantly different between indolent 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)

Using the Wilcoxon test, it was shown that there existed significant differences. The tests were carried out for both hypoxic and oxidative layers for 2 stains; the result for CA9 (**Figure 2C and 2D)**, and for LAMP2b markers **(Figure S2**) as well as architecture, grade, lymphocytes, microcalcifications, and necrosis (**Supplementary Table 1**). As shown in **Figure 2D**, CA9 scoring in hypoxic habitat is much more definitive between indolent and upstaged groups compared to normoxic zone, or the whole duct, or the whole slide scoring as done traditionally (**Figure S1B**). We showed that the CA9-positive cells distribute differently between the 2 patient groups if we focused on hypoxic habitats or oxidative habitats. The performance improvement comparing the count scoring in different habitats and duct scoring in the whole duct suggests that exploring the cell composition and interaction in fine habitats inside ducts are meaningful and necessary.

Defining metabolic niches inside habitats to build spatial machine learning model and improve performance

Previous scoring of hypoxic and normoxic habitats was carried out for each biomarker individually and was restricted to only the count of the positive cells in each habitat. In order to expand the analysis to involve the interaction and relationship between different types of eco-evolutionary marker positive cells, a co-registration step is crucial to create multiplex IHC (mIHC) and map cells onto a common reference 2D space. We chose the HE slides as the reference and registered all IHC slides on to it (**Figure S2**). Note that since our analysis is carried out duct-by-duct, it is not necessary to register the whole slides. Instead, for each duct, we register its IHC staining with the HE staining. This ensures all the downstream analyses could be performed on the same HE coordinates system for each duct. Then we used these mIHC images to define niches of cells representing CA9, LAMP2b, or combination phenotype. We hypothesized that these niches inside habitats using the two phenotypes can be more informative than each marker alone. Thus, we focus on the cell features such as nuclear morphology and texture and cell spatial features inside these niches to explore their effect on upstaging. As illustrated in **Figure 3**, we first map each marker IHC 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, and CA9 and LAMP2b positive niches. We then develop 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.

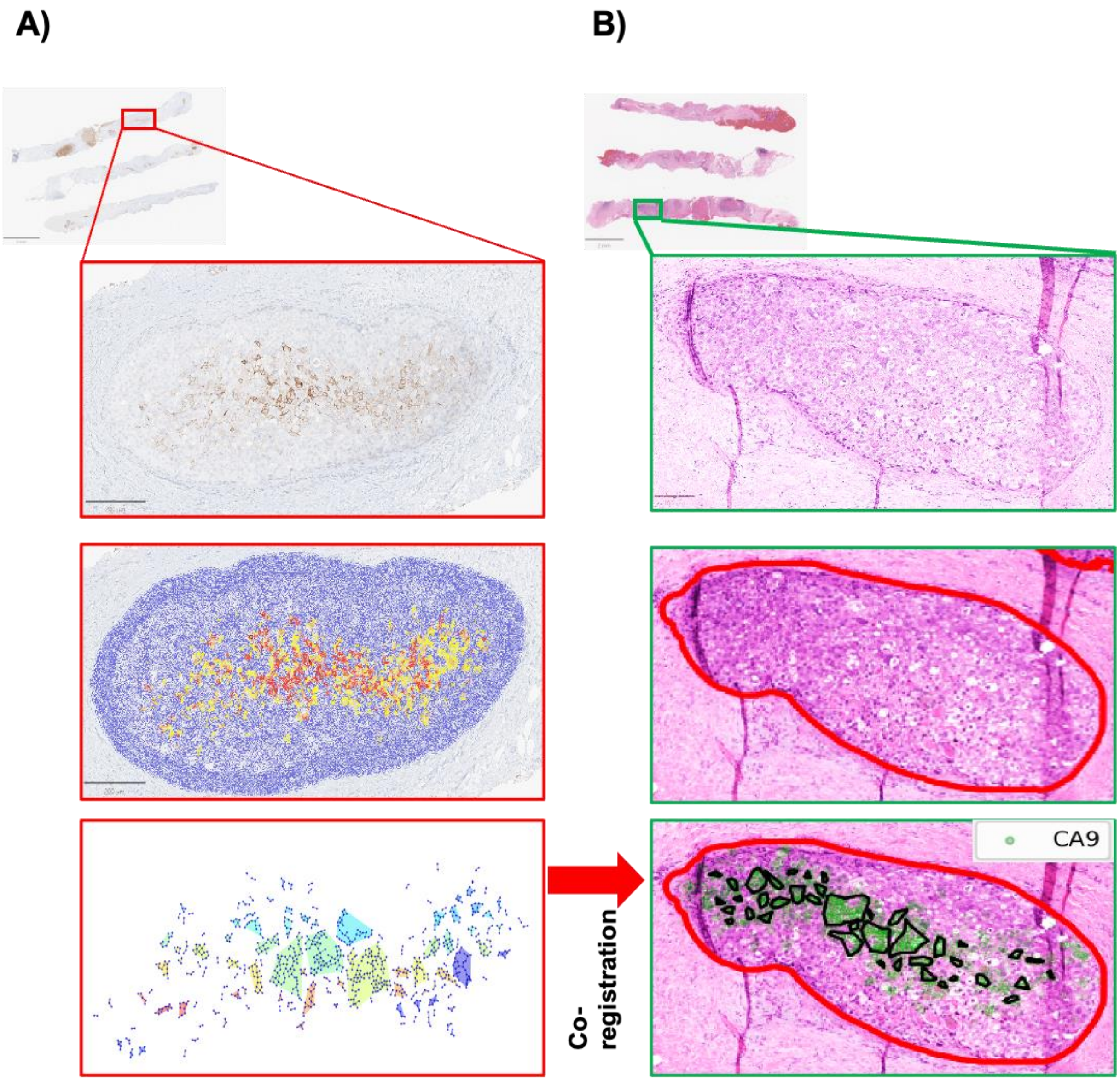

Figure 3. Niches are defined inside habitats from the hypoxia and acidosis markers expression. A) One sample duct from CA9 slide. Top: The original IHC slide. Middle: Cell detection and intensity-based classification using Qupath overlaid on the slide. Bottom: the graph constructed from the CA9 positive 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 slide 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.

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**). Under the habitat level, normoxic and hypoxic zones are analyzed separately. And under the niche level, CA9-positive cells, LAMP2b-positive cells, and CA9- and LAMP2b-positive cells are analyzed separately.

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 biopsy stage. Upon comparing the mean accuracy score and the mean AUC score of all the classifiers, the niche level classifier yielded the best predictive results under both metrics (**Table 1**).

	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.

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

After selecting the best classifier based on the AUC metric, we ran SHAP (SHapley Additive exPlanations) analysis to obtain the SHAP values for each feature, specifically the proportions of each pattern. The most contributing pattern was identified as the one with the maximum SHAP value. Subsequently, a differential analysis was performed for this pattern to identify the top features that significantly differ from other patterns. These features were determined using correlation, mutual information (MI), and maximum relevance minimum redundancy (MRMR) methods. For pattern 5, the common feature set identified included Area_min, Perimeter_min, AreaBbox_min, and F_0< = r <10. A prototype for pattern 5 was then selected, which closely matched the mean values of these top features, and visualized. This process is illustrated in **Figure 4**. By leveraging multi-scale analysis and integrating spatial interactions of CA9 and LAMP2b positive cells through a comprehensive machine learning pipeline, we were able to identify key patterns and features that differentiate between indolent and upstaged DCIS. The use of SHAP analysis and differential analysis allowed us to pinpoint the most impactful patterns and their contributing features, such as Area_min, Perimeter_min, AreaBbox_min, and F_0<=r<10, enhancing the interpretability of the model. Overall, the niche-level analysis yielded the highest accuracy and AUC, underscoring the importance of fine-scale, biologically meaningful regions in predicting clinical outcomes. This approach not only advances our understanding of tumor microenvironments, but also holds promise for more precise prognostic tools in clinical settings.

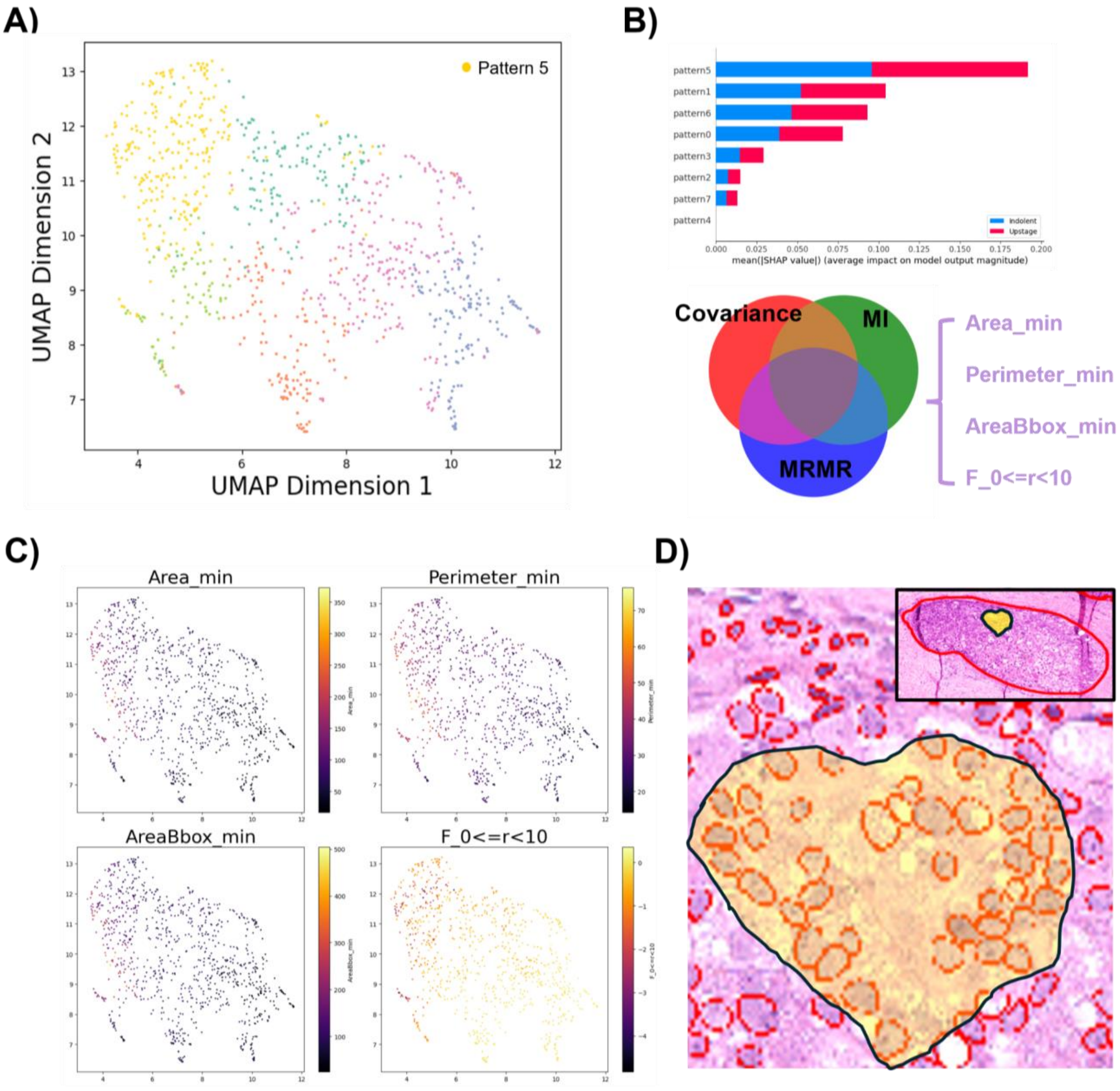

Figure 4. Post Analysis reveals the top contributing patterns and features. A) UMAP of the features of the niches, different colors represent different clusters(patterns) **B)** Top: The impact of each pattern on the classifying result, the proportion of pattern 5 has the greatest impact. Bottom: Using 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.

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 [47]. 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 in order to design the right treatment for breast cancer patients.

Here, we argue that the reason for DCIS overdiagnosis and overtreatment results from conventional frameworks focusing on genetic signature and ignoring phenotypic heterogeneity in tumor ecosystems. Then we interpret complex eco-evolutionary data of cancer cells in their niche using machine learning and pathomics within an innovative ecological and evolutionary dynamic framework. Oxygen habitats are recognized from their variable levels of perfusion and oxygenation. It has been suggested that this variability serves as a significant factor in influencing the ecological diversity, new habitats, and increased tumor heterogeneity leading to diverse evolutionary trajectories. 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 regional selection of cellular phenotypes in distinct subregions. On the other hand, the phenotype of the cells inhabiting these niches can be used to define the habitats. Inhabiting hypoxia, acidosis, and severe nutrient deprivation niches, 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. 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. However, none of these findings were used in an eco-evolutionary designed translational study for biomarker discovery or treatment design. 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 to define habitats that may favor the selection of more aggressive phenotypes, which in turn can be used to predict the upstaging of DCIS.

We curated a retrospective cohort of 84 DCIS patients with histologically confirmed DCIS on core biopsy, followed by surgical excision diagnosed as either DCIS or IDC. We then stained the 2 sequentially sectioned slides for HE, CA9 and LAMP2b and manually annotated 916 single ducts and more than 3000 habitats on all three slides and scored them. 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.

Our study demonstrates the utility of eco-evolutionary principles in understanding DCIS progression. 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.

Method:

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 indolent 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 indolent 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 indolent + 16 upstaged were analyzed. For each sample, we obtained 3 whole slide images, including 1 HE and 2 IHC slides. We conduct 5-fold stratified cross validation, where 4 folds are used for niche clustering and for the training of the indolent/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 used to be stained with primary antibodies of 1:100 dilution of anti-LAMP2 (#ab18529, Abcam), and 1 ug/ml concentration of Goat 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 LAMP2 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 also 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.

MODULE 1: Duct Annotation and Alignment

We annotate and align ducts within all input slides (1 HE + 2 IHCs per sample). For the Bx cohort, there are a small number of ducts per slide. They were annotated by 1 pathologist and trained students. 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.

Manual Annotation of Ducts in the Bx Cohort. QuPath was used as the interface to annotate ducts by the pathologist (Dr. Bai) and the trained students. 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: reactive stroma, oxidative/normoxia, hypoxic/hypoxia, and necrosis. Reactive 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-registrate 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 duct. 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 [30]. For each sample, we chose non-rigid registration and registered the ducts from CA9 and LAMP2b towards the reference HE duct. The co-registration procedure and the qualitative results are shown in **Figure S3 and S4**. 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-ductual 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 an HE duct is done by starDist[25,29] extension on Qupath.

Graph construction for niche detection. After all eco-evo positive cells (i.e., CA9 or LAMP2b positive cells) were annotated and mapped on HE ducts, we were able to perform analysis on an intra-duct level. Since there are still a large amount of eco-evo positive cells within each duct, with diverse spatial context and morphological features, we construct a graph with these cells and detect connected components of the graph as “niches”. Each eco-evo positive cell niche is supposed to have a similar eco-evo phenotype and be spatially coherent. We overlay both CA9 positive and LAMP2b positive cells into the same domain as an approximation of the local eco-evo cell distribution, because the two IHCs are consecutive sections from the same tissue block (**Figure S5**). 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. 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-occurence matrix, contrast, correlation, entropy, homogeneity and intensity .All of these features were calculated following the implementations in the sc-MTOP[35] 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 eco-evo cells of a niche to characterize the niche. For each niche, we identify the concave hull region enclosing its eco-evo positive cells within the HE duct. 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[36] to characterize eco-evo positive cells and their interactions. 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.

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

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,L are calculated for the single-marker cell subsets, and multivariate G_cross, L_cross for different subsets such as CA9-LAMP2b. 'Gest' function and 'Fest' function from 'spatstat' R package were used with Kaplan-Meier estimator[37], and 'Lest' function was used with isotropic correction[38,39].

MODULE 4: Prognostic 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 "indolent" in the future. This establishes the prognostic 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 upstage/indolent status. The prediction power of the classifier sheds light on the prognostic power of the niches and their spatial and cellular features.

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)[42] 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 prognosis. 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 indolent 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.

Resource availability

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). All the staining and annotations are deposited in the physical sciences in oncology network.

Materials availability

This study did not generate new unique reagents.

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. Liliana Davolos and Dr. David Ray for their guidelines on ecological and evolutionary dynamics principles used in the manuscript, as well as for critical reading and providing comments for 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 Y.X, C.C., and M.D. wrote the paper. All authors revised the paper.

Declaration of interests

The authors declare no competing interest.

References:

1. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).

2. Risom, T. *et al.* Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *Cell* **185**, 299–310.e18 (2022).
3. Maley, C. C. *et al.* Classifying the evolutionary and ecological features of neoplasms. *Nat. Rev. Cancer* **17**, 605–619 (2017).
4. Boutry, J. *et al.* The evolution and ecology of benign tumors. *Biochim. Biophys. Acta Rev. Cancer* **1877**, 188643 (2022).
5. Amend, S. R. & 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.* **75**, 2884–2884 (2015).
6. Damaghi, M. *et al.* Collagen production and niche engineering: A novel strategy for cancer cells to survive acidosis in DCIS and evolve. *Evol. Appl.* **13**, 2689–2703 (2020).
7. Lipinski, K. A. *et al.* Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. *Trends Cancer Res.* **2**, 49–63 (2016).
8. Giaquinto, A. N. *et al.* Breast Cancer Statistics, 2022. *CA Cancer J. Clin.* **72**, 524–541 (2022).
9. Strand, S. H. *et al.* Molecular classification and biomarkers of clinical outcome in breast ductal carcinoma in situ: Analysis of TBCRC 038 and RAHBT cohorts. *Cancer Cell* **40**, 1521–1536.e7 (2022).
10. Lehman, C. D. *et al.* National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. *Radiology* **283**, 49–58 (2017).
11. Lips, E. H. *et al.* Genomic analysis defines clonal relationships of ductal carcinoma in situ and recurrent invasive breast cancer. *Nat. Genet.* **54**, 850–860 (2022).
12. Sarhadi, S. *et al.* Omics Integration Analyses Reveal the Early Evolution of Malignancy in Breast Cancer. *Cancers* **12**, (2020).

13. Heselmeyer-Haddad, K. *et al.* Single-cell genetic analysis of ductal carcinoma in situ and invasive breast cancer reveals enormous tumor heterogeneity yet conserved genomic imbalances and gain of MYC during progression. *Am. J. Pathol.* **181**, 1807–1822 (2012).

14. Hanna, W. M. *et al.* Ductal carcinoma in situ of the breast: an update for the pathologist in the era of individualized risk assessment and tailored therapies. *Mod. Pathol.* **32**, 896–915 (2019).

15. Carmeliet, P. & Jain, R. K. Principles and mechanisms of vessel normalization for cancer and other angiogenic diseases. *Nat. Rev. Drug Discov.* **10**, 417–427 (2011).

16. Wu, B. *et al.* Stiff matrix induces exosome secretion to promote tumour growth. *Nat. Cell Biol.* **25**, 415–424 (2023).

17. Persi, E. *et al.* Systems analysis of intracellular pH vulnerabilities for cancer therapy. *Nat. Commun.* **9**, 2997 (2018).

18. Damaghi, M. *et al.* The harsh microenvironment in early breast cancer selects for a Warburg phenotype. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

19. Lobo, R. 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* **4**, 93 (2016).

20. Damaghi, M. *et al.* Chronic acidosis in the tumour microenvironment selects for overexpression of LAMP2 in the plasma membrane. *Nat. Commun.* **6**, 8752 (2015).

21. Ordway, B., Swietach, P., Gillies, R. J. & Damaghi, M. Causes and Consequences of Variable Tumor Cell Metabolism on Heritable Modifications and Tumor Evolution. *Front. Oncol.* **10**, 373 (2020).

22. Gillies, R. J., Verduzco, D. & Gatenby, R. A. Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. *Nat. Rev. Cancer* **12**, 487–493 (2012).

23. Ibrahim-Hashim, A. *et al.* Defining Cancer Subpopulations by Adaptive Strategies Rather

Than Molecular Properties Provides Novel Insights into Intratumoral Evolution. *Cancer Res.* **77**, 2242–2254 (2017).

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* **16**, 1739–1743 (2017).
25. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
26. Freischel, A. R. *et al.* Frequency-dependent interactions determine outcome of competition between two breast cancer cell lines. *Sci. Rep.* **11**, 4908 (2021).
27. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* 234–241 (Springer International Publishing, 2015).
28. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 770–778 (2015).
29. Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. Cell Detection with Star-Convex Polygons. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* 265–273 (Springer International Publishing, 2018).
30. Gatenbee, C. D. *et al.* Virtual alignment of pathology image series for multi-gigapixel whole slide images. *Nat. Commun.* **14**, 4502 (2023).
31. Sugawara, K. Training deep learning models for cell image segmentation with sparse annotations. *bioRxiv* 2023.06.13.544786 (2023) doi:10.1101/2023.06.13.544786.
32. Kirillov, A. *et al.* Segment Anything. *arXiv [cs.CV]* (2023).
33. Ester, M., Kriegel, H., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD* 226–231 (1996).
34. Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. DBSCAN Revisited, Revisited:

Why and How You Should (Still) Use DBSCAN. *ACM Trans. Database Syst.* **42**, 1–21 (2017).

35. Zhao, S. *et al.* Single-cell morphological and topological atlas reveals the ecosystem diversity of human breast cancer. *Nat. Commun.* **14**, 6796 (2023).
36. Baddeley, A., Rubak, E. & Turner, R. *Spatial Point Patterns: Methodology and Applications with R*. (CRC Press, 2015).
37. Baddeley, A. & Gill, R. Kaplan-Meier estimators of interpoint distance distributions for spatial point processes. *IEEE Trans. Inf. Theory* (1993).
38. Ohser, J. On estimators for the reduced second moment measure of point processes. *Series Statistics* **14**, 63–71 (1983).
39. Kendall, W. S. *Stochastic Geometry: Likelihood and Computation*. (Routledge, 2019).
40. Altieri, L., Cocchi, D. & Roli, G. Spatial entropy for biodiversity and environmental data: The R-package SpatEntropy. *Environmental Modelling & Software* **144**, 105149 (2021).
41. Yang, B., Fu, X., Sidiropoulos, N. D. & Hong, M. Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering. in *Proceedings of the 34th International Conference on Machine Learning* (eds. Precup, D. & Teh, Y. W.) vol. 70 3861–3870 (PMLR, 06--11 Aug 2017).
42. 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.* **27**, 1226–1238 (2005).
43. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
44. Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30**, (2017).
45. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

46. Chang, C.-C. & Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, 1–27 (2011).

47. Jardim-Perassi, B. V. *et al.* Multiparametric MRI and Coregistered Histology Identify Tumor Habitats in Breast Cancer Mouse Models. *Cancer Res.* **79**, 3952–3964 (2019).