

## Research Article

# Spatial Transcriptomics Reveal Pitfalls and Opportunities for the Detection of Rare High-Plasticity Breast Cancer Subtypes

Angèle Coutant<sup>a</sup>, Vincent Cockenpot<sup>b</sup>, Lauriane Muller<sup>a</sup>, Cyril Degletagne<sup>a</sup>, Roxane Pommier<sup>a,c</sup>, Laurie Tonon<sup>a,c</sup>, Maude Ardin<sup>a,c</sup>, Marie-Cécile Michallet<sup>a</sup>, Christophe Caux<sup>a</sup>, Marie Laurent<sup>d</sup>, Anne-Pierre Morel<sup>d</sup>, Pierre Saintigny<sup>a,d,e</sup>, Alain Puisieux<sup>d,f,g</sup>, Maria Ouzounova<sup>a,\*</sup>, Pierre Martinez<sup>a,\*</sup>

<sup>a</sup> Université de Lyon, Université Claude Bernard Lyon 1, INSERM 1052, CNRS 5286 Centre Léon Bérard, Cancer Research Center of Lyon, Lyon, France; <sup>b</sup> Department of Pathology, Centre Léon Bérard, Lyon, France; <sup>c</sup> Plateforme de bioinformatique Gilles Thomas, Synergie Lyon Cancer, Centre Léon Bérard, Lyon, France; <sup>d</sup> Centre Léon Bérard, Lyon, France; <sup>e</sup> Department of Medical Oncology, Centre Léon Bérard, Lyon, France; <sup>f</sup> Institut Curie, PSL Research University, Paris, France; <sup>g</sup> Chemical Biology of Cancer Laboratory, CNRS UMR 3666, INSERM U1143, Paris, France

## ARTICLE INFO

## Article history:

Received 14 June 2023

Revised 26 September 2023

Accepted 1 October 2023

## Keywords:

diagnostic markers  
 integrative approaches  
 plasticity  
 rare subtypes  
 spatial transcriptomics

## ABSTRACT

Breast cancer is one of the most prominent types of cancers, in which therapeutic resistance is a major clinical concern. Specific subtypes, such as claudin-low and metaplastic breast carcinoma (MpBC), have been associated with high nongenetic plasticity, which can facilitate resistance. The similarities and differences between these orthogonal subtypes, identified by molecular and histopathological analyses, respectively, remain insufficiently characterized. Furthermore, adequate methods to identify high-plasticity tumors to better anticipate resistance are lacking. Here, we analyzed 11 triple-negative breast tumors, including 3 claudin-low and 4 MpBC, via high-resolution spatial transcriptomics. We combined pathological annotations and deconvolution approaches to precisely identify tumor spots, on which we performed signature enrichment, differential expression, and copy number analyses. We used The Cancer Genome Atlas and Cancer Cell Line Encyclopedia public databases for external validation of expression markers. By focusing our spatial transcriptomic analyses on tumor cells in MpBC samples, we bypassed the negative impact of stromal contamination and identified specific markers that are neither expressed in other breast cancer subtypes nor expressed in stromal cells. Three markers (*BMPER*, *POPDC3*, and *SH3RF3*) were validated in external expression databases encompassing bulk tumor material and stroma-free cell lines. We unveiled that existing bulk expression signatures of high-plasticity breast cancers are relevant in mesenchymal transdifferentiated compartments but can be hindered by abundant stromal cells in tumor samples, negatively impacting their clinical applicability. Spatial transcriptomic analyses constitute powerful tools to identify specific expression markers and could thus enhance diagnosis and clinical care of rare high-plasticity breast cancers.

© 2023 THE AUTHORS. Published by Elsevier Inc. on behalf of the United States & Canadian Academy of Pathology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

\* Corresponding authors.

E-mail addresses: [maria.ouzounova@lyon.unicancer.fr](mailto:maria.ouzounova@lyon.unicancer.fr) (M. Ouzounova), [pierre.martinez@lyon.unicancer.fr](mailto:pierre.martinez@lyon.unicancer.fr) (P. Martinez).



## Introduction

Breast cancer is one of the most prominent types of cancers,<sup>1</sup> with 2.3 million women diagnosed in 2020 (Globocan, 2020 International Agency for Research on Cancer). Patient stratification relies on the presence of specific targetable alterations in the estrogen receptor, progesterone receptor, and *HER2* genes. These genetic properties are generally well recapitulated by the broad PAM50 transcriptomic signatures: luminal A, luminal B, *HER2*-enriched, basal-like, and normal-like.<sup>2</sup> Clinical approaches typically rely on the immunohistochemical expression of those 3 markers, conducted routinely during the pathological examination of breast cancer specimens. Triple-negative breast cancer (TNBC) patients, however, lack any of the targetable estrogen receptor, progesterone receptor, and *HER2* alterations, leading to scarce therapeutic options and low survival rates despite promising results with antibody-drug conjugates.<sup>3,4</sup> Rarer subtypes, such as claudin-low (CL) tumors or metaplastic breast carcinoma (MpBC), have furthermore been associated with high plasticity, a cellular property facilitating dynamic phenotypic changes and the subsequent emergence of nongenetic therapeutic resistance.<sup>5</sup> Anticipating the ability of breast cancer cells to adapt via plasticity is thus of paramount importance for effective therapeutic targeting. However, the driving mechanisms of tumor plasticity remain poorly understood, and no standard method exists to accurately detect or quantify it for patient stratification.

High-plasticity breast cancer subtype identification typically relies on either molecular or histopathological analyses. CL tumors were originally defined by transcriptomic analyses, with a phenotype similar to basal cells lacking expression of claudins 3, 4, and 7, and other cell-cell adhesion markers.<sup>6</sup> They represent 3% to 5% of all breast cancers<sup>7,8</sup> and are generally associated with strong stemness features. The evolutionary trajectories underlying their malignant progression are, however, still debated. CLs typically display high expression of epithelial-mesenchymal transition (EMT) factors,<sup>6</sup> known to foster phenotypic plasticity and stemness,<sup>9,10</sup> and enhance tolerance to oncogenic stress, thereby mitigating genomic instability.<sup>11</sup> Recent work further suggests the existence of different CL classes: CL1, CL2, and CL3.<sup>7</sup> CL1 tumors are believed to arise directly from malignantly transformed mammary stem cells (MaSC). They appear to be the most representative of previous observations of EMT-driven plasticity and genomic stability, displaying the highest intrinsic expression of EMT markers and the lowest fraction of genome altered (FGA).

MpBC are a heterogeneous group histopathologically defined by the presence of a nonepithelial tumor component believed to occur through transdifferentiation.<sup>12</sup> MpBCs, accounting for 0.2% to 2% of all invasive breast carcinomas, (IBCs) are usually triple-negative and often associated with poorer survival rates.<sup>13</sup> Different subtypes exist according to the transdifferentiated component,<sup>13</sup> including but not restricted to spindle, chondroid, and osseous cells.<sup>13,14</sup> MpBC can present  $\geq 1$  metaplastic compartments, which are frequently admixed with a component of IBC of no special type (NST). MpBC diagnosis remains challenging, and adequate markers are lacking to correctly classify specific subtypes within this highly heterogeneous disease.<sup>15</sup> Similar to CLs, many of these subtypes display high EMT marker expression and resemble mammary tumor-initiating stem cell-like cells based on transcriptomic data.<sup>16</sup> Although CL tumors frequently undergo metaplastic differentiation,<sup>6</sup> a thorough comparison between these subtypes, defined by different approaches, has never been carried out.

Here, we aimed at further characterizing these plasticity-associated breast cancer subtypes, defined either molecularly or histopathologically, via spatial transcriptomics (SpaT). Based on previously described<sup>7,8</sup> genomic instability and CL-associated bulk gene signatures, we identified 3 putative CL tumors (CL-like) and 4 non-CL, genomically unstable TNBC samples as controls. We also selected 4 MpBCs through histopathology (2 spindle cell, 1 chondroid, and 1 IBC-NST compartment from a mixed spindle cell tumor) and performed SpaT analyses on all 11 samples. We report that, unlike unstable TNBC samples, the tumor compartment of CL-like tumors did not recapitulate expression patterns expected from bulk analyses. We demonstrate that existing CL expression signatures are significantly upregulated in MpBCs with mesenchymal transdifferentiation but that the prevalence of stromal cells can hinder clinical applicability and lead to false-positive diagnoses. This pitfall highlights the need to integrate histopathological approaches into transcriptomic analyses to define more robust signatures that are specific to high-plasticity tumor cells.

By focusing SpaT expression analyses on tumor cells in MpBC samples, we identified specific expression markers that are neither expressed in other subtypes nor expressed in stromal cells. SpaT-derived markers could thus enhance the diagnosis and clinical care of rare high-plasticity breast cancers in the future.

## Materials and Methods

### Low and High Genomic Instability Sample Identification

We analyzed 87 fresh frozen TNBC samples with paired whole exome sequencing (WES) data collected at the Centre Léon Bérard from 379 samples with RNA sequencing (RNA-seq) data assembled as part of the MyPROBE project (17-RHUS-0008). Only samples with more than 30% tumor purity (estimated percentage of tumor cells), as estimated by FACETS,<sup>17</sup> had been included in this data set. We used the already processed RNA-seq and WES data available for all samples. Using the precalculated FGA, we identified 3 samples with low genomic instability (FGA, <10%) and 6 samples with high instability (FGA, >75%). Samples CLB-17, CLB-52, and CLB-74 were considered CL-like, and samples CLB-14, CLB-23, CLB-37, and CLB-51 were considered unstable TNBC (Supplementary Table S1). An additional CL-like sample (CLB-11 or CL-like 4) was discarded after in-depth investigation and identification of an erroneously low FGA (actual FGA, >10%). However, a tumor-free slide from this discarded tumor was included in the controls.

### Spatial Transcriptomic Analyses

Visium Spatial 3' v1 slides for fresh frozen samples (10x Genomics) were processed according to the manufacturer's guidelines. Briefly, tissue was fixed on each capture area with methanol for 30 minutes at  $-20^{\circ}\text{C}$ , stained with hematoxylin and eosin, imaged using the 3DHISTECH Panoramic Scan II scanner, permeabilized for 12 minutes, and reverse-transcribed to create RNA libraries tagged with unique molecular identifiers (UMIs). RNA libraries were then sequenced in 2 batches on a NovaSeq 6000 sequencer (Illumina), targeting 50,000 reads/spot. Each Visium "slide" comprised 4 capture areas; each capture area could be used to analyze a single slide of frozen material from a given sample (ie, 1 Visium slide corresponds to the analysis of 4 tissue slides, each in a separate capture area). Each tissue slide analyzed by Visium was

given a unique identifier starting with M. The first batch included tissue slides M1 to M11 (CL-like, unstable TNBC, and normal controls), and the second batch included tissue slides M13 to M16 (metaplastic TNBCs). Correspondence between patient and slide identifiers is listed in [Supplementary Table S1](#). The Space Ranger software (10x Genomics) was used to process the raw data. Stereoscope<sup>18</sup> was used for deconvolution analyses using only the TNBC data from a single-cell breast cancer atlas<sup>19</sup> and excluding plasmablasts that were initially found to be overrepresented in deconvolution results (included cell types: cancer epithelial, normal epithelial, endothelial, cancer-associated fibroblasts, T cell, B cell, myeloid, and perivascular-like cells).

Signature enrichment analyses were performed using the AUCell tool,<sup>20</sup> designed for UMI-based single-cell data that present similar biases to those of our SpaT data. To maximize the capture of weakly expressed genes in all signatures, we used the maximum threshold advised by AUCell developers (0.2). This provided higher minimal scores compared to lower thresholds (0.5, 0.1, and 0.15) without altering the overall observations ([Supplementary Fig. S1](#)).

#### *Combining Histopathological Annotations and Spatial Transcriptomics on Fresh Frozen Tissue*

We analyzed a total of 15 Visium capture areas, including 3 controls extracted from tumor samples with epithelial compartments but no identifiable tumor on the slide. All slides were annotated by a breast pathologist (V.C.) to separate tumor/nontumor and epithelial/nonepithelial compartments as well as to eliminate spots corresponding to folded tissue and artifacts. The SpaT data and paired annotations have been publicly deposited as a series on the Gene Expression Omnibus (GSE213688).

Tumor spots were first annotated by the pathologist. To later investigate expression markers that would not be biased by the prevalence of mesenchymal stromal cells, we restricted the analysis of stromal spots to the following pathologist annotations, exclusively: fibrosis, fibrosis (peritumoral), fibrous stroma, tumor stroma, and tumor stroma fibrous, the last 2 separated on visual appreciation of the cellular density and collagen abundance in the stroma. Tumor and stromal spots identified by the pathologist were further refined using scores obtained by RNA-based computational deconvolution. This ensured that only the most enriched spots in the population of interest were analyzed by combining human expertise and computational data mining. For each spot, the stereoscope software reports individual scores for each of the 8 cell types on which it was trained: the higher the score, the higher (likelier) the proportion of the related cell type in a spot, and the sum of these 8 scores is always 1. Tumor spots with a cancer epithelial score <0.1, as attributed by the stereoscope software, were discarded. In addition, spots annotated as tumor by the pathologist in the MpBC\_chondroid sample with unusual B cell stereoscope scores >0.1 were also discarded. For spots annotated as stromal by the pathologist, those with either normal epithelial or cancer epithelial scores >0.1 in the stereoscope deconvolution output were discarded.

#### *Copy Number Alteration Profiles*

Three slides that were devoid of tumor cells upon histopathological examination were included as normal references: M1 and M9 (CL like3) and M7 (CL like4). The expression profiles of epithelial spots present on these slides provided a more relevant

reference than those of other cells, in which cell type-specific expression patterns local to genome segments could hamper copy number alteration (CNA) identification. Individual spot copy number (CN) profiles were produced with the infercnvPlus tool (<https://github.com/CharleneZ95/infercnvPlus>, based on inferCNV of the Trinity CTAT project: <https://github.com/broadinstitute/inferCNV>). For each slide, all spots annotated as either tumoral or epithelial were pooled using epithelial spots as references.

This produced a relative CN measure per gene per spot, which we compiled into a major cytoband per spot matrix, by computing the average CN per cytoband. Sample-specific profiles were then obtained by averaging the relative CN measures per cytoband across all spots of each given sample. Comparable profiles were obtained for bulk samples, by averaging per cytoband the segmented logR values reported by FACETS<sup>17</sup> analysis, from the WES data of CL-like and unstable samples.

Gain/loss profiles were calculated for each bulk sample by attributing values of 1 (gain), 0 (normal), or -1 (loss) to each cytoband, according to whether their mean logR exceeded the median logR of the sample  $\pm 0.6$ , as implemented by Van Loo et al.<sup>21</sup> A threshold was used to calculate similar gain/loss profiles in Visium samples, according to whether the cytoband CNA score exceeded the threshold. Its value was optimized to minimize the pairwise distance across all cytobands between the bulk-derived and SpaT-derived gain/loss profiles for the 4 unstable tumors, which provided the most reliable and informative data.

CN profiles were calculated for all Molecular Taxonomy of Breast Cancer International Consortium IntClusters (METABRIC) using the difference between gains and losses (number of gains minus number of losses) reported for each gene in each cluster and then summed per cytoband as for our other CN profiles. Correlation-based distances were obtained by calculating the Pearson R correlation between 2 samples, minus 1, and then divided by -2. This gives a minimum distance of 0 for perfectly correlated samples and a maximum distance of 1 for perfectly anticorrelated samples.

#### *Differential Expression Analyses on Spatial Transcriptomics Data*

All tumor samples were first individually normalized using the SCTransform function of the Seurat R package.<sup>22</sup> All 11 samples were then merged together and renormalized using the SCTransform function, resulting in a matrix of 22,058 genes by 14,905 spots. The MAST R package<sup>23</sup> was used to perform differential expression analyses on SpaT data. Fold changes (FC) were defined as the mean expression in the population of interest divided by the mean expression in the control population.

The CL, unstable control, and normal samples (M1-M11) were distributed across 3 Visium slides and were sequenced together in the first batch. All 4 MpBC samples (M13-M16) were on the same Visium slide and were sequenced in the second separate batch. We thus adapted our design to remove potential biases stemming from batch effects and different sequencing depths. The Uniform Manifold Approximation and Projection of all MpBC tumor spots was obtained without spots from other samples, to prevent batch effects, and then normalized using the harmony software<sup>24</sup> to account for technical effects between capture areas. For the same reasons, we initially defined differentially expressed (DE) genes only among the metastatic samples. We used the tumor spots from the MpBC\_NST sample as a nonmetastatic reference and the spots annotated as fibrosis and tumor stroma from the MpBC\_spindle2 sample as a stromal reference. To account for the uneven number of tumor spots between the 2 spindle MpBC samples, we first performed DE analysis between them and

retained all genes with a nonsignificant difference in expression ( $P > .05$ ; absolute  $\log_2$  FC,  $< 2$ ). We then performed DE analyses between combined tumor spots from both spindle samples compared to tumor spots from the NST sample and the stromal references separately. Genes upregulated in the spindle tumor spots were defined as those having a  $\log_2$  FC of  $>1.5$  and a corrected  $P$  value of  $<.001$  (Benjamini-Hochberg). Genes overexpressed in MpBC tumor spots in both analyses (vs non-MpBC tumor and stroma) were considered candidate MpBC markers. We further validated their relevance in the tumor and stromal spots from the CL-like and unstable samples using the same method but with a higher  $\log_2$  FC threshold ( $>2$ ) in this larger data set. Only genes significantly overexpressed in MpBC samples in both analyses (non-MpBC tumor and stroma) were considered internally validated markers. For external validation, we used the clinical information, including sample metaplastic status, and messenger RNA expression (z-normalized RNA-seq log RSEM) data from The Cancer Genome Atlas (TCGA) breast cancer Firehose Legacy cohort retrieved through the cBioPortal.<sup>25,26</sup> Cancer Cell Line Encyclopedia (CCLE) data, z-normalized relative to diploid samples, were also obtained from the cBioPortal. CL classification of the CCLE breast cancer cell lines was retrieved from the studies by Pommier et al<sup>7</sup> and Prodhomme et al.<sup>27</sup> Differential expression was assessed using Wilcoxon rank-sum tests for each gene of interest.

**Results**

*Identification of Claudin-Low–Like, Unstable, and Metaplastic Breast Tumors*

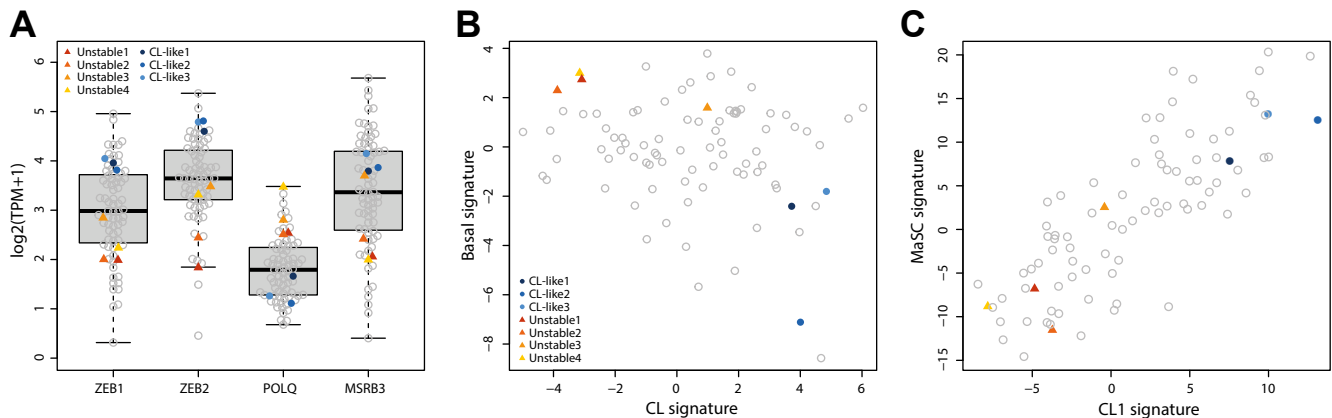
We focused on identifying CL tumors displaying high intrinsic plasticity and low genomic instability, as described by the CL1 subclassification.<sup>7</sup> We analyzed 379 TNBC samples from the Centre Léon Bérard for which RNA-seq data were available, 87 of which also had paired whole exome data. To initially identify potential CL1 tumors based on genomic stability, we selected 3 tumors with FGA  $<10\%$ . For control TNBC samples, we focused on those least resembling CL1 tumors and identified 6 genomically unstable tumors with FGA  $>75\%$  (Supplementary Table S1). The low-FGA tumors displayed RNA expression patterns highly concordant with CL, and particularly CL1, phenotypes: high *ZEB1*, *ZEB2*, and *MSRB3* and low *POLQ*<sup>11,27</sup> (Fig. 1A).

We thus considered these 3 samples as highly relevant CL-like candidates. Of the 6 high-FGA samples, to be used as controls, we selected the 4 high-FGA samples displaying the most contrasting patterns regarding these genes (Supplementary Fig. S2). We performed PAM50 centroid classification on the 379 samples from our cohort merged with the TCGA BRCA data set. This revealed that the 3 CL-like samples were similar to luminal A, and the 4 unstable control samples were closer to the Basal subtype. Furthermore, these findings were consistent with gene set enrichment analyses of established CL and Basal-like gene signatures from the study by Prat et al<sup>6</sup> and the CL1 and MaSC signatures from the study by Pommier et al<sup>7</sup> (Fig. 1B, C, Supplementary Figs. S3–S5). These tumors were histologically reviewed to confirm their histotype; all were classified as TNBC-NST according to the *WHO Classification of Tumours of the Breast, Fifth edition*, criteria, and no metaplastic component was identified in these samples.

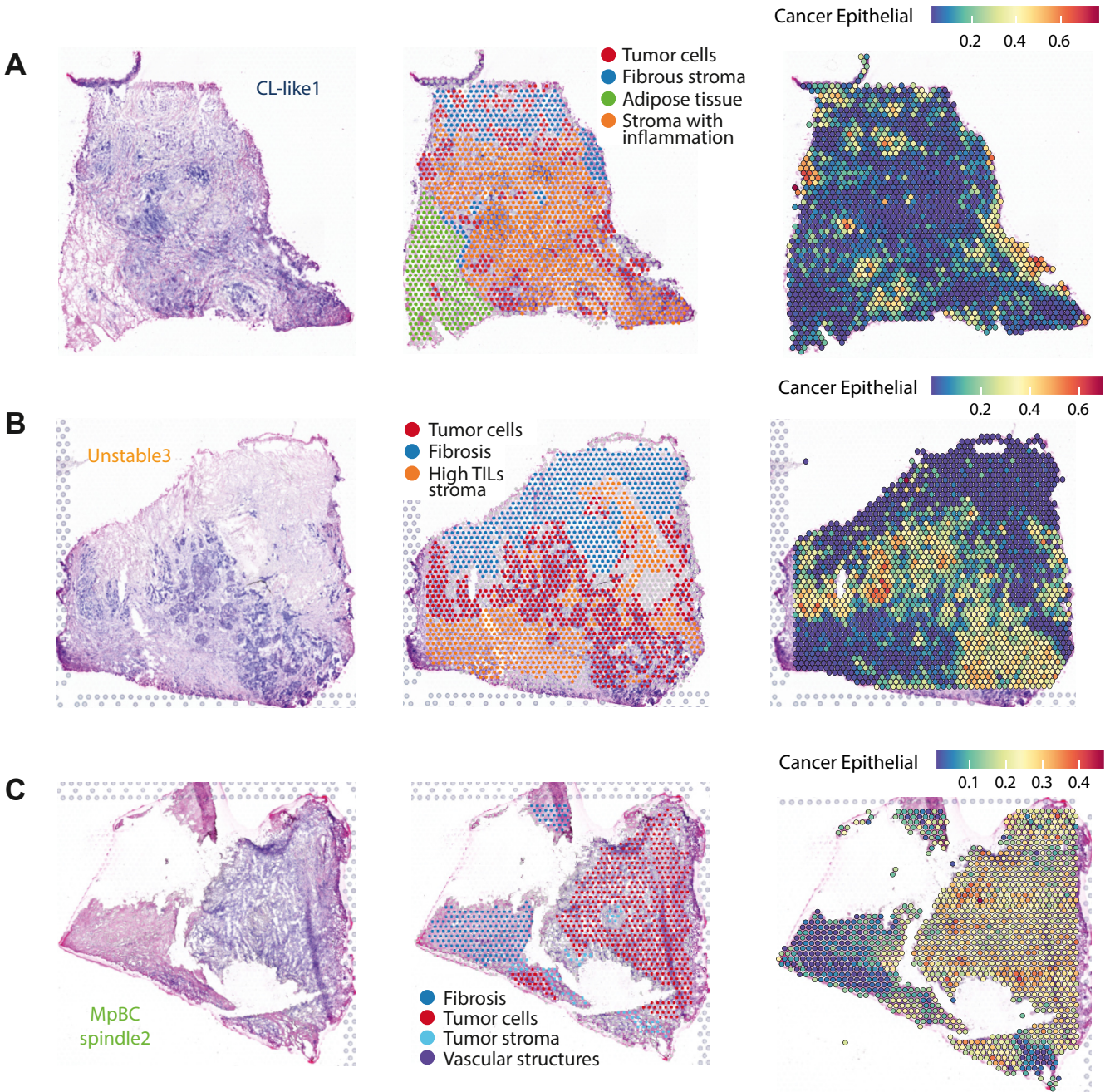
To broaden the analysis of high-plasticity breast tumor types, we further selected 4 MpBC of different subtypes: 2 spindle cell carcinomas (MpBC\_spindle), 1 carcinoma with pure chondroid differentiation (MpBC\_chondroid), and 1 IBC-NST compartment from a tumor diagnosed as a mixed spindle cell and IBC-NST (MpBC\_NST). Their subtypes were initially determined during routine histopathological evaluation for clinical diagnosis on formalin-fixed, paraffin-embedded samples and were then reviewed by a breast pathologist. Only tumors for which fresh frozen material was available were selected in order to use the same SpaT technology as for CL and control TNBCs. Frozen specimens were also reviewed on hematoxylin and eosin–stained slides to assess the sampled components and select optimal samples to be used for SpaT.

*Spatial Transcriptomic Data Deconvolution Reflects Histopathological Annotations*

We performed SpaT analyses on a total of 14 slides comprising 3 CL-like tumors, 4 unstable TNBCs, 4 MpBCs, and 3 adjacent normal tissue with no detectable tumor cells. We uncovered medians of 85% reads under tissue, 2398 genes per spot, and 5083 UMIs per spot (Supplementary Table S1). Estimation of spot cellular composition using the stereoscope deconvolution software confirmed that spots histopathologically identified as tumor



**Figure 1.** Identification of claudin-low (CL)-like and unstable samples. (A) Expression of CL1 markers *ZEB1*, *ZEB2*, *POLQ*, and *MSRB3* in the 87 MyPROBE triple-negative breast cancer samples with paired whole exome sequencing and RNA sequencing data for 3 low-fraction of genome altered samples considered to be CL-like and 4 high-fraction of genome altered samples considered to be unstable. Scores for gene set enrichment analyses across the 87 samples for (B) the CL and Basal expression signatures and (C) the CL type 1 (CL1) and mammary stem cell (MaSC) expression signatures.



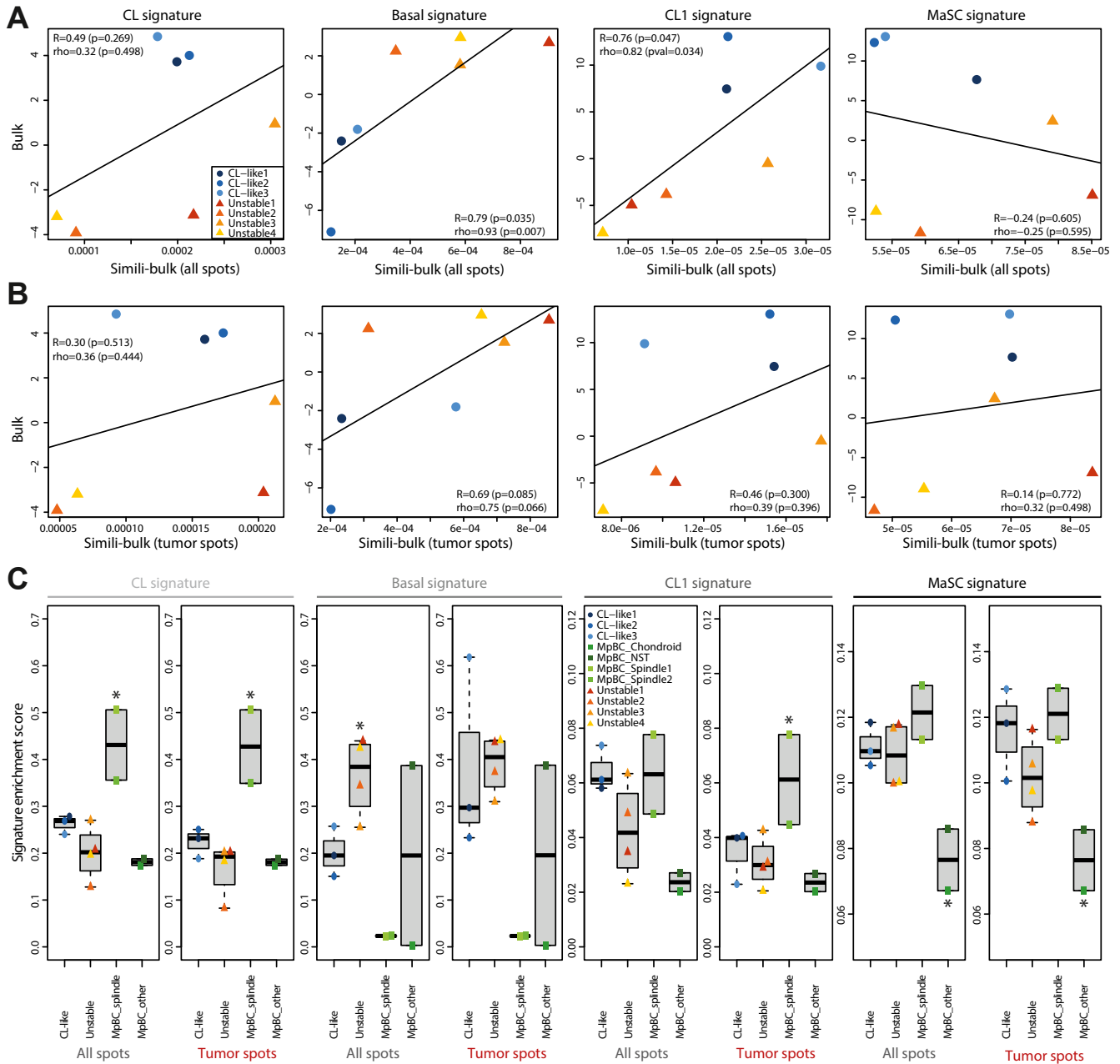
**Figure 2.**

Overlaying pathology annotations and in silico deconvolution in spatial transcriptomics data. Hematoxylin and eosin staining (left), pathologist annotations (center), and per-spot deconvolution-based cancer epithelial score (right) for samples (A) CL-like1, (B) unstable3, and (C) MpBC\_spindle1. Unlabeled gray spots were either considered to be artifacts or could not be annotated with confidence by the pathologist. CL, claudin-low; MpBC, metaplastic breast carcinoma; TIL, tumor-infiltrating lymphocyte.

displayed more tumor-associated expression patterns ( $P < .001$ , *t* test; Fig. 2, Supplementary Figs. S6-S10) and that adjacent normal slides were tumor-free (Supplementary Fig. S9). This confirmed that the RNA profiles from tumor-labeled spots reflected pathological annotations.

Computational deconvolution was based on individual RNA profiles of cells extracted from tumors of epithelial origin<sup>19</sup> and could identify cancer cells in all but 1 of the 4 MpBC slides: the MpBC\_spindle1 sample (Supplementary Fig. S8), in which the strong mesenchymal nature of metaplastic tumor cells led to spots being predicted as containing mostly cancer-associated

fibroblasts. In later analyses, we used both pathological annotations and thresholds on computer-derived deconvolution scores to identify genuine tumor and nontumor spots (see Materials and Methods). Of note, all SpaT analyses were performed on tumor sections that were distinct from the ones initially used for bulk analyses, which can lead to sampling bias. Importantly, we report that the tumor slides from the CL-like2 and CL-like3 analyzed by SpaT were annotated as noninvasive in situ carcinoma, while both patients were diagnosed with invasive TNBC. This may imply phenotypic differences compared to the sections included in bulk analyses initially performed on these samples.



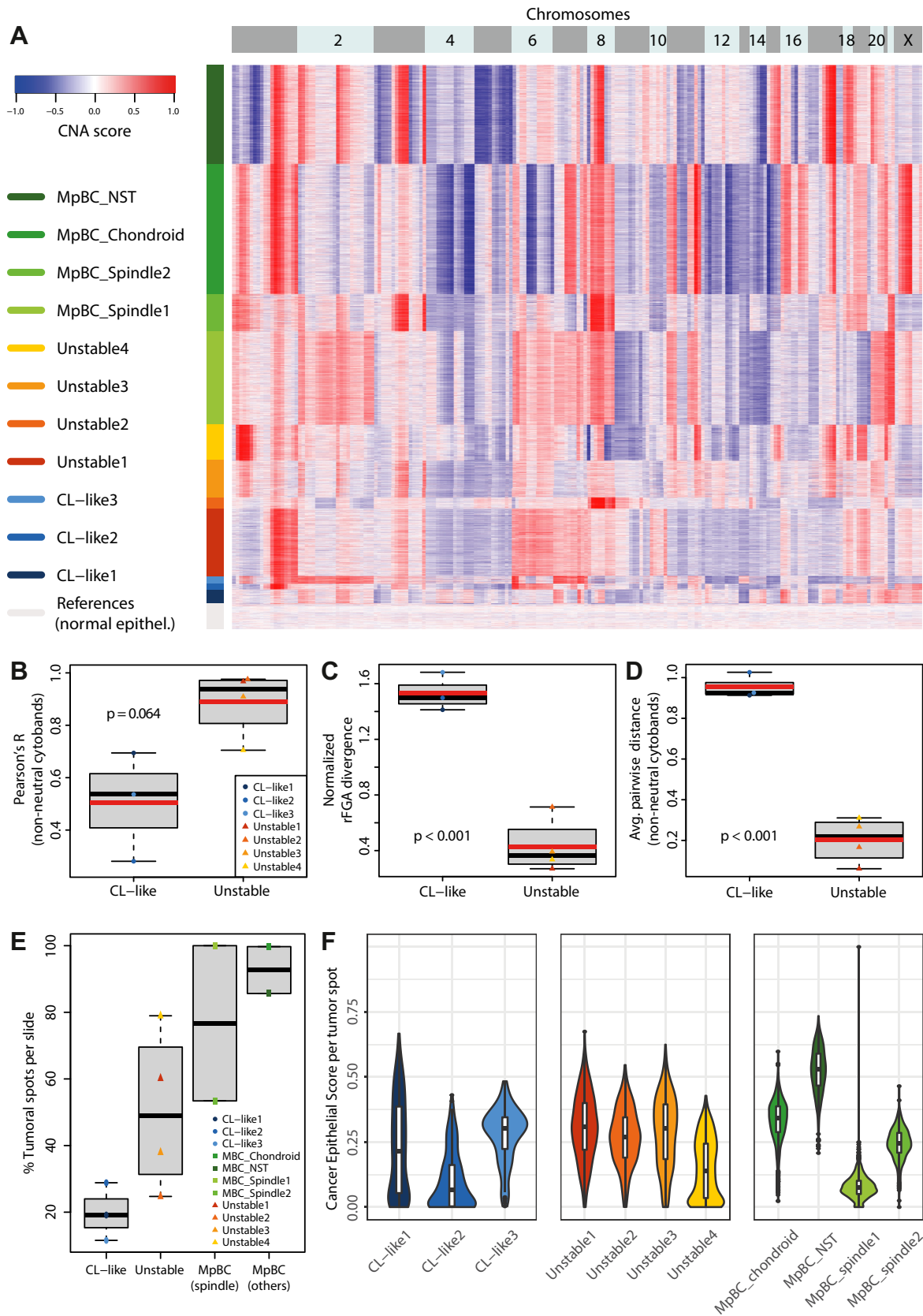
**Figure 3.**

Spatialized enrichment analyses of existing plasticity-associated expression signatures. (A) Correlation for gene set enrichment scores between bulk- and spatial transcriptomics-derived pseudobulk data for claudin-low (CL), Basal, CL1, and mammary stem cell (MaSC) expression signatures using all annotated spots. (B) Correlation for gene set enrichment scores between bulk- and spatial transcriptomics-derived pseudobulk data for CL, Basal, CL1, and MaSC expression signatures focusing on tumor spots. (C) Average per-spot gene set enrichment scores for the CL, Basal, CL1, and MaSC expression signatures in each sample of CL-like, unstable, and metaplastic breast carcinoma (MpBC) sample types. For clarity, the MpBC samples were dichotomized according to the presence or absence of a spindle cell transdifferentiated compartment on the captured area ("spindle" or "other," respectively). The asterisks highlight groups of samples whose signature enrichment scores were significantly different from those of all other pooled samples ( $P < .05$ , Wilcoxon rank-sum tests). NST, no special type

*Spatialization Sheds Light on the Impact of the Microenvironment on Molecular Plasticity Signatures*

To assess how SpaT recapitulates the bulk signal used to classify the CL-like and unstable control samples, we generated pseudobulk data by pooling all spots on each slide. In addition, histopathological annotations of tumor spots allowed us to investigate the prominence of CL, Basal, CL1, and MaSC signatures

on the entire surface of slides as well as in tumor spots. When all spots per slide were pooled, correlations between the bulk and pseudobulk (all spots) data were statistically significant for the Basal and CL1 signatures ( $\rho = 0.93$ ,  $P = .007$ ;  $\rho = 0.82$ ,  $P = .034$ , respectively) but not for the CL and MaSC signatures (Fig. 3A). However, no significant correlation was found when focusing on tumor spots (Fig. 3B). This suggests that SpaT data were less similar to bulk data when focusing on tumor spots.



**Figure 4.**

Unstable spatial transcriptomics (SpaT)-derived copy number (CN) profiles are less divergent from their bulk counterpart. (A) Copy number alteration (CNA) profile calculated for each tumor spot of all 11 cancer samples. This score represents a per-spot relative measure, with values close to  $-1$  (blue) corresponding to the highest loss of genomic material and values close to  $1$  to the highest gain. For each spot (y axis), values were averaged per major cytoband (x axis), that is, the average of the scores obtained for each

We further analyzed each signature using either all spots or only tumor spots in all samples (Fig. 3C). The original CL signature was significantly higher in spindle cell MpBC samples, particularly in tumor spots (Fig. 3C;  $P = .036$ , Wilcoxon rank-sum tests). The CL1 signature was also enriched in spindle cell MpBC tumor spots ( $P = .036$ ). The 3 spindle cell and chondroid MpBC samples also displayed a significantly lower enrichment of the Basal signature when grouped together ( $P = .012$ , both cases; Fig. 3C). This suggests that these existing signatures are relevant for the identification of breast cancer transdifferentiation, particularly to a mesenchymal phenotype.

The restriction of signature enrichment analyses to tumor spots had little impact on unstable control samples. In CL-like samples, however, CL and CL1 signatures decreased in this case, while the Basal signature increased, albeit only with borderline significance due to limited data points ( $P = .169$ ,  $P = .020$ , and  $P = .254$ , respectively,  $t$  tests). This further indicates that the specific transcriptomic profiles of tumor cells in these samples did not reflect the bulk-based signal used to identify CL-like samples at diagnosis.

#### *Claudin-Low-Like Samples Poorly Recapitulate Bulk Copy Number Alterations*

We determined relative CN profiles per major cytoband for each SpaT tumor spot using the infercnvPlus software. Intra-tumor heterogeneity was very low in the regionally restricted SpaT data, with per-spot CN profiles displaying little variability within a sample but being markedly dissimilar across samples (Fig. 4A). As such, spot-specific profiles were then averaged to obtain per-sample CN profiles (Supplementary Fig. S11A). These profiles derived from measured RNA quantities are only relevant as relative values within each sample and, thus, cannot be used to determine absolute CNs. We thus defined a threshold to identify regions of chromosomal loss and gain based on the unstable control samples, which provided more reliable estimates in terms of number of CNAs per sample and signal quality (Supplementary Figs. S11B, C and S12-S14; see Materials and Methods). This allowed us to calculate relative fractions of genome altered (rFGA) and investigate distances and correlations between bulk and SpaT CN profiles (Fig. 4B, C). Unlike bulk data, CL-like samples did not appear genomically stable, and their SpaT-derived rFGAs were not significantly different from those of unstable control samples ( $P = .88$ ,  $t$  test). However, we found that CL-like CN profiles were less correlated with their bulk counterparts than unstable profiles, both on the entire genome (Supplementary Fig. S15) and exclusively on regions of chromosomal gain or loss ( $P = .064$  and  $P = .074$ , respectively,  $t$  test). Analyses of differences in rFGA and average pairwise distances between bulk and SpaT profiles (see Materials and Methods) further confirmed that SpaT data in unstable control samples, rather than in CL-like samples, better recapitulated the bulk CNA information (Fig. 4C, D, both  $P < .001$ ,  $t$  test).

We further analyzed the similarities between CN profiles from all 6 CL-like bulk and SpaT samples (Supplementary Fig. S16A) and compared them to the 10 IntClusters from Curtis et al.<sup>28</sup> (Supplementary Fig. S16B). We used distances based on Pearson R correlation to evaluate the similarities between per-cytoband CN profiles (minimum 0 for perfect correlation; maximum 1 for perfect anticorrelation; see Materials and Methods). This first revealed that the bulk CN profiles of the CL-like1 and CL-like3 samples were closest to their SpaT counterpart than to any other sample (and vice versa). Despite the low quality of the bulk signal, this suggests that the real CN profile was detectable in the bulk, albeit greatly diluted by a large fraction of normal diploid stromal cells for these 2 samples. In addition, CL-like CN profiles were most similar to IntClusters 4 (CL-like1) and 10 (CL-like2 and CL-like3), substantiating our previous report<sup>7</sup> stating that IntClust4 comprised all highly genetically stable samples from the CL1 subtype, and IntClust10 was enriched in the CL3 subtype, more similar to genetically unstable basal TNBC samples.

#### *Higher Stromal Content in Claudin-Low-Like Samples*

We also observed that SpaT slides from CL-like tumors contained a significantly smaller fraction of tumor spots (Fig. 4E;  $P = .02$ , Wilcoxon rank-sum test) and a lower fraction of cancer epithelial cells per spot than the other tumors, as estimated by the stereoscope, even when including transdifferentiated MpBC samples (Fig. 4F;  $P < .001$ , Wilcoxon rank-sum test). Given that tumor spots in CL-like samples did not display a strong signal for the highly mesenchymal CL expression signatures, this strongly suggests that an insufficient percentage of tumor cells in CL-like bulk samples could have both artificially increased enrichment scores for CL expression signatures and hampered the detection of genuine CNAs.

#### *Spatial Transcriptomics Can Identify Tumor-Specific Markers That Are Robust Against High Stromal Content*

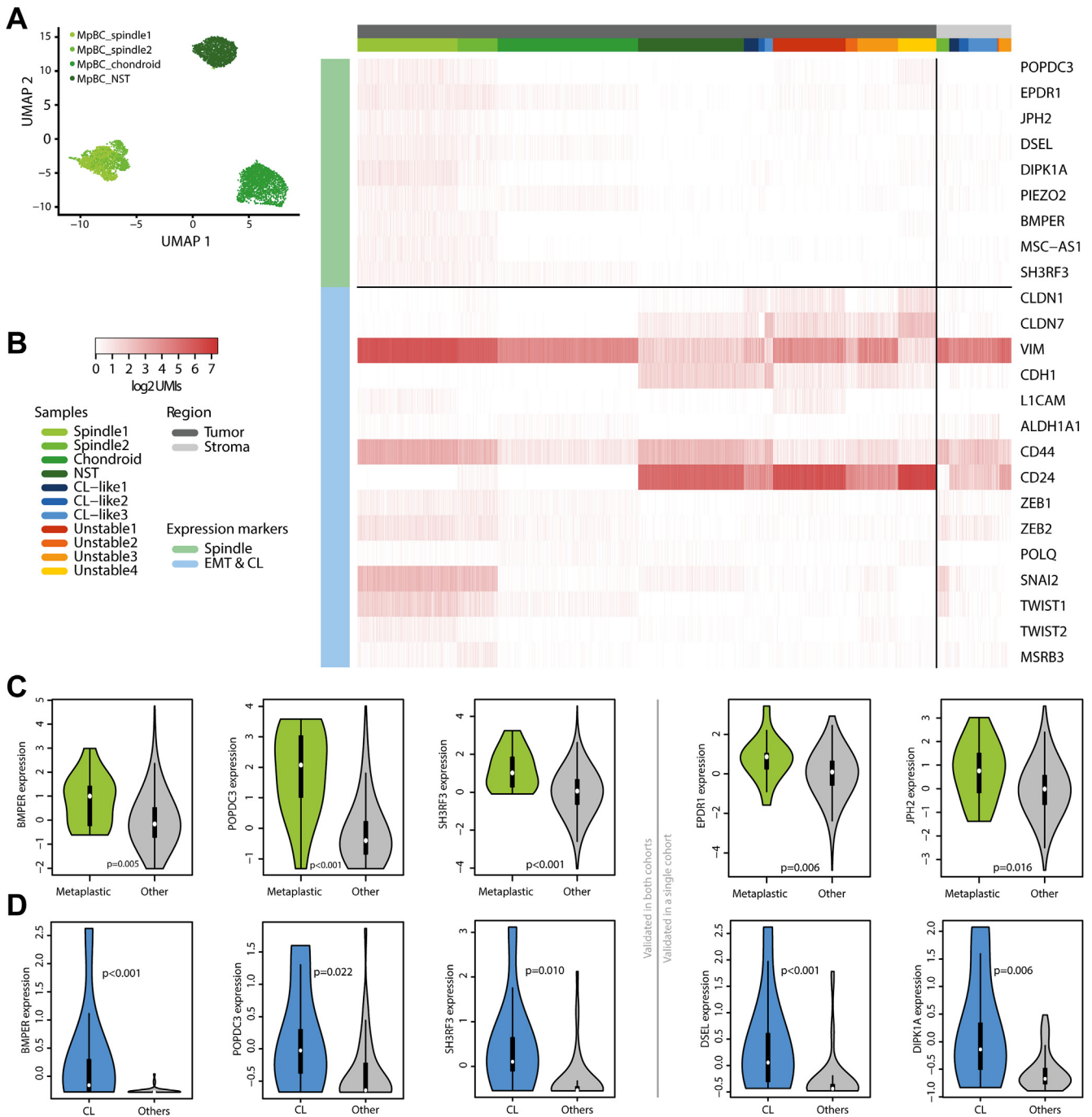
Clustering analyses on the tumor spots from MpBC samples revealed that spindle cell samples clustered together, highlighting that they shared common, spindle cell-specific expression patterns (Fig. 5A). We thus harnessed SpaT data to identify genes overexpressed in MpBC tumor cells compared to both tumor cells of NST and healthy stromal cells.

With 1983 spots per slide on average (range, 1055–3037), our SpaT data provided a very powerful basis for differential expression analyses, even with few samples. We then aimed at mitigating the impact of the low number of samples and maximizing reproducibility (see Materials and Methods). Based on the high number of spots available, we split our cohort into internal discovery (MpBC samples) and validation (CL-like and unstable samples) sets to prevent overfitting (Table). We identified genes significantly overexpressed in spindle cell or chondroid spots

---

individual gene in the cytoband. Samples for which more tumor spots were identified take more space on the y axis. (B) Pearson correlation between bulk logR and SpaT-derived average CNA scores, averaged per major cytoband, in claudin-low (CL)-like and unstable samples. Only cytobands harboring nonneutral CNAs in the SpaT data were included in the correlation calculation. (C) Normalized divergence between bulk- and SpaT-derived relative fractions of genome altered (rFGAs) in CL-like and unstable samples. (D) Average pairwise distances between bulk- and SpaT-derived CN profiles in CL-like and unstable samples. CN profiles were based on a  $-1$  (loss),  $0$  (neutral), or  $1$  (gain) attributed to each cytoband. Only the cytobands reported as nonneutral in the SpaT profiles were included in the distance calculation. (E) Percentage of tumor spots in each captured area, per sample type. (F) Cancer epithelial score, as reported by stereoscope-based deconvolution, for all tumor spots in each sample. MpBC, metaplastic breast carcinoma; NST, no special type.





**Figure 5.**

Spatial transcriptomics-derived expression markers. (A) Uniform Manifold Approximation and Projection (UMAP) of all tumor spots from the 4 metaplastic breast carcinoma (MpBC) samples analyzed by spatial transcriptomics. (B) Expression heatmap of the newly identified spindle cell MpBC markers (top, light green) and known epithelial-mesenchymal transition (EMT) and claudin-low (CL) markers (bottom, light blue). Expression is reported in log<sub>2</sub> number of unique molecular identifiers per spot, ranging from white (no expression) to red (high expression), and was analyzed in both tumor spots (left, black) and stromal spots (right, gray). (C) Expression of validated spindle cell MpBC genes *BMPER*, *EPDR1*, *JPH2*, *POPDC3*, and *SH3RF3* in an external cohort of 1108 samples analyzed per bulk RNA sequencing and stratified by tumor metaplastic status (14 metaplastic tumors in total). (D) Expression of validated spindle cell MpBC genes *BMPER*, *POPDC3*, *SH3RF3*, *DSEL*, and *DIPK1A* in an external cohort of 51 cancer cell lines analyzed per RNA sequencing and stratified by CL status (9 CL samples in total). NST, no special type.

compared to both NST and stromal spots from MpBC samples, combining 2 separate differential expression analyses. We then selected those that were overexpressed compared to both tumor spots and stromal spots from CL-like and unstable control samples.

Using this design, we identified subtype-specific genes with low expression in nonmetaplastic tumor cells and stromal cells. Nine genes were overexpressed in the 2 spindle cell MpBC samples: *BMPER*, *DIPK1A*, *DSEL*, *EPDR1*, *JPH2*, *PIEZO2*, *POPDC3*, *MSC-AS1*, and *SH3RF3* (Fig. 5B). We also investigated known EMT markers,<sup>6-8</sup>

**Table**  
Number of spots used in each design for differential expression analyses

Type of spots	No. of spots in cohort		
	Discovery (spindle)	Discovery (chondroid)	Validation
Metaplastic breast carcinoma	1977	1984	NA
Non-metaplastic breast carcinoma	1492	1492	2719
Stroma	181	181	875

Metaplastic breast carcinoma spots were compared to both non-metaplastic breast carcinoma spots and stromal spots separately, first in the discovery cohort and then in the validation cohort. NA, not available.

none of which were significant in our analyses, and whose expression patterns were less specific to spindle cell MpBC tumor cells. The same analysis conducted on the MpBC\_chondroid samples identified 8 genes whose overexpression was specific to chondroid MpBC cells: *HSPB6*, *VGLL3*, *PTX3*, *GFRA1*, *MT1G*, *MT1E*, *CHI3L2*, and *SAAI* (Supplementary Fig. S17).

However, these genes were identified from a small number of samples ( $n = 2$  for spindle cell;  $n = 1$  for chondroid). To assess their relevance in larger cohorts, we investigated their expression in 2 complementary external data sets: the TCGA breast cancer expression data set comprising 1108 samples,<sup>29</sup> including 14 metaplastic samples, and the 51 stroma-free breast cancer cell lines from the CCLE,<sup>30</sup> including 9 samples classified as CL. Long non-coding RNA *MSC-AS1* was the only gene for which expression data were unavailable in both sets. In the TCGA data set, samples were classified as non-CL ( $n = 940$ ), CL1 ( $n = 69$ ), CL2 ( $n = 42$ ), or CL3 ( $n = 57$ ) according to previous work based on expression signatures.<sup>27</sup> *BMPER*, *EPDR1*, *JPH2*, *POPDC3*, and *SH3RF3* were significantly overexpressed in metaplastic samples (for all,  $P < .05$ , Wilcoxon rank-sum tests; Fig. 5C, Supplementary Fig. S18). In the CCLE data set, *BMPER*, *DIPK1A*, *DSEL*, *POPDC3*, and *SH3RF3* were significantly overexpressed in CL cell lines (Fig. 5D, Supplementary Fig. S19). All significant  $P$  values held after Benjamini-Hochberg correction in each data set.

These results validate that 3 of the 8 analyzable stroma-independent spindle cell MpBC markers (*BMPER*, *POPDC3*, and *SH3RF3*) were highly relevant in metaplastic tumors and CL cell lines from 2 independent external cohorts. Four more genes could, furthermore, be validated in 1 external cohort (*EPDR1* and *JPH2* in the TCGA data; *DIPK1A* and *DSEL* in the CCLE data) but not the other. These results suggest that leveraging SpaT data to focus solely on tumor cells can yield powerful expression markers that are less likely to be biased by stromal cell prevalence. This could help identify high-plasticity tumors more reliably and improve molecular classification.

In addition, all genes except *PIEZO2* displayed a significant overexpression in CL subtypes (Supplementary Fig. S20) in the TCGA data, confirming the similarities between CL and MpBC tumors. The overlap between the 2 subtypes showed significant enrichment as 57% of TCGA (8/14) MpBCs were classified as CL (4 CL1, 1 CL2, and 3 CL3), which represent only 15% of tumors ( $P < .001$ , Fisher exact test). As for chondroid markers, *VGLL3* and *PTX3* could both be validated in MpBC samples in the TCGA data due to their significant overexpression after multiple testing (Supplementary Fig. S21).

## Discussion

Rare breast cancer subtypes, associated with cellular plasticity features, remain difficult to diagnose and treat. Two subtypes in

particular, molecularly defined CL tumors and MpBC typically defined by histopathology, appear to overlap. Here, based on recent advances in SpaT, we shed light on the similarities and discrepancies between these plasticity-associated breast cancer subtypes and assess the opportunities and pitfalls their routine classification presents. We identified 3 putative CL samples and 4 unstable TNBCs using gene expression and CN data as well as 4 MpBCs reviewed by a breast pathologist (V.C.). We analyzed a single slide of each of the 11 samples by SpaT along with 3 slides from adjacent normal breast tissue with no detectable tumor material as controls.

We investigated 4 previously reported signatures (CL vs Basal from Prat et al.,<sup>6</sup> CL type 1, and MaSCs from Pommier et al.<sup>7</sup>) and found they were highly relevant in MpBC, particularly for the spindle cell subclassification. However, CL-like samples, defined by molecular analyses, were heavily impacted by stromal cell prevalence, hampering the detection of genuine CNAs and biasing RNA signals toward mesenchymal gene overexpression, which is a feature of CL tumors. Restricting our SpaT analyses to spots harboring tumor cells revealed that these cells did not recapitulate the low-FGA, high-CL expression features that prompted our initial assessment of these samples as putative CL based on bulk analyses.

Overall, these findings illustrate the obstacles hindering molecular-based identification of high-plasticity subtypes as overrepresentation of stromal cells in a sample can lead to false positivity. This is likely the case for the 3 tumors we identified as putative CL tumors, which may have led to their PAM50 classification as luminal rather than basal samples by diluting the basal expression signal. It is clear that at least a subset of tumors molecularly classified as CL would also be histologically defined as spindle cell MpBCs. However, the nature of CL tumors that are not spindle cell MpBCs remains insufficiently characterized. It will be important to determine whether these samples merely display strong stromal prevalence or whether they reflect an additional specific tumoral phenotype. Indeed, if such cells, displaying CL transcriptomic features but without being histologically identified as metaplastic, exist, our results here highlight that SpaT analyses could both validate their existence and uncover more specific markers. This would help refine the classification of rare breast cancers and more clearly define nonoverlapping bona fide subtypes with specific clinical outcomes. Despite all 3 CL-like tumors being reported as invasive carcinoma, 2 of the 3 slides cut from frozen leftover material were annotated as in situ carcinoma during our SpaT analyses. Our findings may thus not be representative of all CL tumors.

Here, we could further harness the spatial information to analyze differential expression between the different compartments of tumor samples at near single-cell resolution. Using internal and external validation procedures, we could identify *BMPER*, *POPDC3*, and *SH3RF3* as robust spindle cell MpBC markers, whose detection is unlikely to be affected by the stromal content of samples. Of note, *BMPER* was reported to promote invasive phenotypes and angiogenesis in cancer,<sup>31</sup> and *SH3RF3* was shown to promote stem-like properties in breast cancer.<sup>32</sup> We also identified *MSC-AS1*, a long noncoding RNA that has been linked to both osteogenic differentiation<sup>33</sup> and oncogenesis,<sup>34</sup> as an additional potential spindle cell MpBC marker. However, we could not validate its relevance due to its absence in the external data set.

Our SpaT analyses revealed that although breast tumor cells displaying CL-like mesenchymal properties could be detected in spindle cell MpBC samples, their classification using existing molecular signatures in bulk samples remains error-prone. However, despite our limited sample size, SpaT proved extremely powerful for the identification of genes that are highly specific to

transdifferentiated tumor cells. This is of particular interest for translational research on rare subtypes as large cohorts are difficult to obtain. SpaT analysis can prove very powerful to identify candidate markers with no previous expectations on a limited number of samples before performing targeted and more affordable validation studies on a larger scale. In rare breast cancers, larger SpaT studies than ours could provide more specific expression signatures for MpBC and CL tumor cells. This may prove useful to help diagnose such complex cases, for which integrative histologic and molecular approaches are required to overcome their respective limitations.

#### Acknowledgments

The biological samples used in this study were prepared by the Biological Resource Center of the Centre Léon Bérard (BB-0033-00050), Lyon, France.

#### Author Contributions

P.S., A.-P.M., A.P., M.O., and P.M. designed the study. M.L., M.-C.M., and C.C. collected the samples. V.C., L.M., C.D., and M.O. processed the samples. V.C. annotated the slides. A.C., R.P., L.T., M.A., and P.M. performed computational analyses. A.C., V.C., M.O., and P.M. wrote the manuscript.

#### Data Availability

Spatial transcriptomics data have been deposited as a series on the Gene Expression Omnibus database: GSE213688. The RNA and whole exome sequencing data from the MyPROBE cohort, analyzed to identify the claudin-low-like and unstable samples in this study, have neither been published nor released elsewhere. The identifiers will, however, be consistent with those used here, and data will be accessible upon request to the authors.

#### Funding

This research was financially supported by the Institut Thématique Multi-Organismes (ITMO) Cancer of AVIESAN (Alliance Nationale pour les Sciences de la Vie et de la Santé, National Alliance for Life Sciences & Health) within the framework of the Cancer Plan, as well as by grants RHU MyPROBE and ANR-17-RHUS-0008. This research was additionally supported by SIRIC LYriCAN (INCa-DGOS-Inserm\_12563). A.C. is funded by a scholarship from the Ligue Nationale contre le Cancer.

#### Declaration of Competing Interest

The authors declare that they have no conflict of interests.

#### Ethics Approval and Consent to Participate

The Biological Resource Center (BRC) of the Centre Léon Bérard (no. BB-0033-00050), biological material collection and storage facility, is registered at the Ministry of Research (DC-2008-99 and AC-2019-3426). Samples were used in the context of patient diagnosis. This study was approved by the ethical review board of the Centre Léon Bérard. The BRC of the Centre Léon Bérard has an

AFNOR NFS96900 (no. 2009/35884.2) and ISO 9001 (certification no. 2013/56348.2) quality certificates for clinical trials, ensuring scientific rigor for sample conservation, traceability, and quality, as well as ethical rules observance, and defined rules for transferring samples for research purposes (Ministry of Health for activities authorization nos. AC-2019-3426 and DC-2008-99). Sample usage is reviewed by a multidisciplinary committee before any transfer. The samples are properly coded to ensure anonymity of the donors or usage of any clinical information. The material used in the study was collected in agreement with all applicable laws, rules, and requests of French and European government authorities, including the patient's informed consent.

#### Supplementary Material

The online version contains supplementary material available at <https://doi.org/10.1016/j.labinv.2023.100258>

#### References

- Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021;71(3):209–249. <https://doi.org/10.3322/caac.21660>
- Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol.* 2009;27(8):1160–1167. <https://doi.org/10.1200/JCO.2008.18.1370>
- Bardia A, Hurvitz SA, Tolaney SM, et al. Sacituzumab govitecan in metastatic triple-negative breast cancer. *N Engl J Med.* 2021;384(16):1529–1541. <https://doi.org/10.1056/NEJMoa2028485>
- Schmid P, Cortes J, Dent R, et al. Event-free survival with pembrolizumab in early triple-negative breast cancer. *N Engl J Med.* 2022;386(6):556–567. <https://doi.org/10.1056/NEJMoa2112651>
- Black JRM, McGranahan N. Genetic and non-genetic clonal diversity in cancer evolution. *Nat Rev Cancer.* 2021;21(6):379–392. <https://doi.org/10.1038/S41568-021-00336-2>
- Prat A, Parker JS, Karginova O, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.* 2010;12(5):R68. <https://doi.org/10.1186/bcr2635>
- Pommier RM, Sanlaville A, Tonon L, et al. Comprehensive characterization of claudin-low breast tumors reflects the impact of the cell-of-origin on cancer evolution. *Nat Commun.* 2020;11(1):3431. <https://doi.org/10.1038/s41467-020-17249-7>
- Fougner C, Bergholtz H, Norum JH, Sørbye T. Re-definition of claudin-low as a breast cancer phenotype. *Nat Commun.* 2020;11(1):1787. <https://doi.org/10.1038/s41467-020-15574-5>
- Mani SA, Guo W, Liao MJ, et al. The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell.* 2008;133(4):704–715. <https://doi.org/10.1016/j.cell.2008.03.027>
- Morel APP, Lièvre M, Thomas C, Hinkal G, Ansieau S, Puisieux A. Generation of breast cancer stem cells through epithelial-mesenchymal transition. *PLoS One.* 2008;3(8):e2888. <https://doi.org/10.1371/journal.pone.0002888>
- Morel AP, Ginestier C, Pommier RM, et al. A stemness-related ZEB1–MSRB3 axis governs cellular pliancy and breast cancer genome stability. *Nat Med.* 2017;23(5):568–578. <https://doi.org/10.1038/nm.4323>
- Reddy TP, Rosato RR, Li X, Moulder S, Pivnicka-Worms H, Chang JC. A comprehensive overview of metaplastic breast cancer: clinical features and molecular aberrations. *Breast Cancer Res.* 2020;22(1):121. <https://doi.org/10.1186/S13058-020-01353-Z>
- Rakha EA, Allison KH, Ellis IO, Horii R, Masuda S, Penault-Llorca F. *WHO Classification of Tumours of the Breast.* International Agency for Research on Cancer; 2019. Accessed July 25, 2022. [https://books.google.com/books/about/WHO\\_Classification\\_of\\_Tumours\\_of\\_the\\_Bre.html?id=n-hMAQAACAJ](https://books.google.com/books/about/WHO_Classification_of_Tumours_of_the_Bre.html?id=n-hMAQAACAJ)
- McCart Reed AE, Kalaw E, Nones K, et al. Phenotypic and molecular dissection of metaplastic breast cancer and the prognostic implications. *J Pathol.* 2019;247(2):214–227. <https://doi.org/10.1002/PATH.5184>
- Khoury T. Metaplastic breast carcinoma revisited; subtypes determine outcomes: comprehensive pathologic, clinical, and molecular review. *Surg Pathol Clin.* 2022;15(1):159–174. <https://doi.org/10.1016/J.PATH.2021.11.011>
- Hennessy BT, Gonzalez-Angulo AM, Stemke-Hale K, et al. Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics. *Cancer Res.* 2009;69(10):4116–4124. <https://doi.org/10.1158/0008-5472.CAN-08-3441>
- Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* 2016;44(16):e131. <https://doi.org/10.1093/NAR/GKW520>

18. Andersson A, Bergenstråhle J, Asp M, et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun Biol*. 2020;3(1):565. <https://doi.org/10.1038/s42003-020-01247-y>
19. Wu SZ, Al-Eryani G, Roden DL, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet*. 2021;53(9):1334–1347. <https://doi.org/10.1038/s41588-021-00911-1>
20. Aibar S, González-Blas CB, Moerman T, et al. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods*. 2017;14(11):1083–1086. <https://doi.org/10.1038/nmeth.4463>
21. Van Loo P, Nordgard SH, Lingjærde OC, et al. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A*. 2010;107(39):16910–16915. <https://doi.org/10.1073/pnas.1009843107>
22. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*. 2015;33(5):495–502. <https://doi.org/10.1038/NBT.3192>
23. Finak G, McDavid A, Yajima M, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol*. 2015;16:278. <https://doi.org/10.1186/s13059-015-0844-5>
24. Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*. 2019;16(12):1289–1296. <https://doi.org/10.1038/s41592-019-0619-0>
25. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2(5):401–404. <https://doi.org/10.1158/2159-8290.CD-12-0095>
26. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*. 2013;6(269):p11. <https://doi.org/10.1126/SCISIGNAL.2004088>
27. Prodhomme MK, Pommier RM, Franchet C, et al. EMT transcription factor ZEB1 represses the mutagenic POLθ-mediated end-joining pathway in breast cancers. *Cancer Res*. 2021;81(6):1595–1606. <https://doi.org/10.1158/0008-5472.CAN-20-2626>
28. Curtis C, Shah SP, Chin SF, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012;486(7403):346–352. <https://doi.org/10.1038/nature10983>
29. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490(7418):61–70. <https://doi.org/10.1038/nature11412>
30. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483(7391):603–607. <https://doi.org/10.1038/nature11003>
31. Heinke J, Kerber M, Rahner S, et al. Bone morphogenetic protein modulator BMPER is highly expressed in malignant tumors and controls invasive cell behavior. *Oncogene*. 2012;31(24):2919–2930. <https://doi.org/10.1038/ONC.2011.473>
32. Zhang P, Liu Y, Lian C, et al. SH3RF3 promotes breast cancer stem-like properties via JNK activation and PTX3 upregulation. *Nat Commun*. 2020;11(1):2487. <https://doi.org/10.1038/S41467-020-16051-9>
33. Zhang N, Hu X, He S, et al. LncRNA MSC-AS1 promotes osteogenic differentiation and alleviates osteoporosis through sponging microRNA-140-5p to upregulate BMP2. *Biochem Biophys Res Commun*. 2019;519(4):790–796. <https://doi.org/10.1016/j.bbrc.2019.09.058>
34. Cao C, Zhong Q, Lu L, et al. Long noncoding RNA MSC-AS1 promotes hepatocellular carcinoma oncogenesis via inducing the expression of phosphoglycerate kinase 1. *Cancer Med*. 2020;9(14):5174–5184. <https://doi.org/10.1002/CAM4.3080>