nature communications

Article

https://doi.org/10.1038/s41467-026-70242-4

A hormetic transcriptional program coregulates invasion, proliferation and dormancy to define metastatic potential

Received: 24 July 2025

Accepted: 23 February 2026

Published online: 04 March 2026

Check for updates

Raúl Jiménez-Castaño[1], Nitin Narwade[1], Gema Moreno-Bueno[2,3,4,5], Berta Sánchez-Laorden[1], Joan Galcerán[1], Khalil Kass Youssef[1] & M. Angela Nieto[1]

Metastasis remains the leading cause of cancer mortality, yet the factors that determine metastatic competence and the timing of its acquisition are not well understood. Using human breast cancer cohorts and multi-omics in preclinical models, here we show that cells exit the primary tumor with pre-set metastatic potential. Integrating spatial transcriptomics with single-cell gene expression and chromatin profiles, we identify the transcription factor Prrx1 as a master regulator of dissemination: beyond its known role in invasion, it represses proliferation by acting on cyclins and cell-cycle inhibitors (Ccnd1/2, Cdkn2a/b/c) and activates a dormancy program (Gas6, Mme, Ogn). Intermediate Prrx1 levels optimize the trade-off between invasion and proliferation/dormancy, producing a hormetic (nonlinear) relationship between Prrx1 expression and metastatic burden. Combined invasion and proliferation signatures strongly stratify breast cancer prognosis, underscoring the clinical relevance of this phenotypic interplay. These findings indicate that metastatic competence emerges from the integration of Prrx1 co-regulated molecular programs, explaining why some invasive cells enter dormancy while others resume growth.

Despite significant advances in cancer treatment, which have extended median survival rates, most patients ultimately succumb to metastatic relapse[1]. Metastasis is a highly complex, multi-step process that only a small fraction of cancer cells manages to complete[2]. The timing of metastasis is highly variable; some patients are diagnosed with metastatic disease at the time of detection, while in others, metastases appear years after initial treatment[3–5]. Recent advances have shed light on the biology of metastatic cells, highlighting cell-autonomous characteristics and microenvironmental effectors as key determinants of metastatic success[6]. Additionally, there is little evidence supporting a strong genetic basis for metastasis[7–9], suggesting that cancer cell heterogeneity and plasticity may underlie metastatic behavior[10]. However, the biological features that define metastatic cells and when they are acquired remain poorly understood.

One possibility is that pre-existing cell states influence the future behavior of metastatic cells, due to pure intrinsic characteristics or distinct responsiveness to environmental signals at the metastatic site[11]. Epithelial-mesenchymal transition (EMT) is a well-known driver of cell plasticity and tumor dissemination[12,13]. EMT is a dynamic process in which cells can lie at a spectrum of intermediate states between the epithelial and mesenchymal phenotypes[14]. Different EMT states in cancer cells have been associated with varying metastatic potential, with hybrid E/M cells often exhibiting the highest metastatic competence[15–18].

[1]Instituto de Neurociencias (CSIC-UMH), Sant Joan d'Alacant, Spain. [2]Instituto de Investigaciones Biomédicas "Sols-Morreale" CSIC-UAM, Madrid, Spain. [3]CIBERONC Centro de Investigación Biomédica en Red de Cancer, ISCIII, Madrid, Spain. [4]MD Anderson Cancer Center International Foundation, Madrid, Spain. [5]Translational Cancer Research Group, Area 3 Cancer, Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid, Spain. e-mail: anieto@umh.es

Recently, we described an EMT-driven intratumor heterogeneity in breast cancer, identifying two distinct trajectories, respectively associated with invasion or inflammation, each controlled by different EMT transcription factors[19]. We found that among them, Prrx1 is required for invasion and, consequently, metastasis. However, Prrx1 high expression prevents metastatic colonization[20,21]. These contrasting results suggest that the integration of different cellular programs that shape metastatic potential may be influenced by Prrx1 levels.

By integrating patient data, single-cell sequencing modalities, genetic mouse models bearing tumors with distinct levels of *Prrx1* and engineered cell lines, we demonstrate that *Prrx1* expression in the primary tumor dictates metastatic competence by simultaneously regulating key cellular traits: invasion, proliferation, and dormancy. In addition, we found that *Prrx1* levels exhibit a hormetic behavior, a biphasic dose–response in which intermediate levels promote metastasis, while high levels have an inhibitory effect[22,23], reconciling the apparently conflicting data mentioned above[19–21]. Furthermore, we identify Prrx1 as a master regulator that integrates invasion, proliferation, and dormancy programs and show that a combinatorial approach using molecular signatures derived from our analyses enables the identification of a breast cancer subgroup with poor prognosis.

Results

Non-linear correlation between PRRX1 levels and metastatic incidence in breast cancer patients

We had previously shown that high levels of *Prrx1* expression in the primary tumor correlated with the lack of metastasis in breast cancer patients[20]. To better assess the relationship between PRRX1 expression in cancer cells and metastatic incidence without the influence of stromal-derived PRRX1 expression, we analyzed 94 (grade 3) infiltrating ductal carcinoma (IDC) tumors in an adjuvant setting from patients who underwent surgery between 2010 and 2011. PRRX1 was visualized using immunohistochemistry in combination with an anti-keratin antibody cocktail to specifically identify cancer cells, and intratumor heterogeneity and intensity of PRRX1 signal in cancer cells were evaluated by a pathologist. Based on both the intensity and heterogeneity of PRRX1 expression in cancer cells, tumors were classified into three groups: the majority (82%, with no detectable PRRX1-positive cancer cells) as PRRX1-negative; 10% as PRRX1-intermediate, characterized by a sparse presence of cells expressing intermediate PRRX1 levels; and the remaining 8% as PRRX1-high, containing a substantial number of cancer cells strongly positive for PRRX1 (Fig. 1a). To substantiate the existence of these three groups in a quantitative manner, a subset of samples containing all groups were analyzed using image-analysis software on a single-cell basis (see "Methods"). Unsupervised clustering of the patients' scores based on (i) percentage of PRRX1+ and (ii) intensity of PRRX1 in cancer cells confirmed the three patient groups and individual assignment (Fig. 1b). PRRX1-negative tumors were predominantly classified as Luminal A or B (65%), with a smaller fraction corresponding to HER2 (12%) or Basal (23%). This is compatible with previous observations showing that *Prrx1*-deficient PyMT tumors are highly differentiated[19]. In contrast, tumors with PRRX1-intermediate levels were more evenly distributed among Luminal B (22%), HER2 (33%), and Basal (45%), and all PRRX1-high tumors belonged to the Basal subtype.

To better understand the pattern of PRRX1 expression in human breast cancer, we analyzed the METABRIC database[5,24,25]. Interestingly, almost 20% of patients show amplification of the *PRRX1* locus (Supplementary Fig. 1a). Analysis of the Integrative Clusters described in the original METABRIC manuscripts revealed a clear enrichment of IntClust8 in patients bearing amplifications (Supplementary Fig. 1b), which is expected because this subgroup bears amplification of 1q chromosomal regions where PRRX1 is located. However, IntClust4 showed even higher *PRRX1* expression than IntClust8 (Supplementary Fig. 1c) and is considered to have a good prognosis[24]. Regarding molecular subtypes, Claudin-low tumors exhibited the highest PRRX1 expression (Supplementary Fig. 1d), consistent with the strong EMT activation observed in this subtype[26,27]. Most Claudin-low tumors fall within IntClust4 (Supplementary Fig. 1e), the cluster with the highest PRRX1 levels (Supplementary Fig. 1f).

Finally, we observed a non-linear correlation between PRRX1 levels and metastatic incidence. The PRRX1-intermediate group exhibited the highest metastatic burden, with around 78% of patients developing metastases, compared to 31% in the PRRX1-negative group and only 12% in the PRRX1-high group (Fig. 1d and Supplementary Table 4). These findings are aligned with previous observations in patients showing that low Prrx1 expression is associated with worse survival[20,21]. These results suggest that PRRX1 levels, rather than its mere presence or absence, may be a critical determinant of the metastatic potential of cancer cells.

In vivo genetic modeling of different *Prrx1* levels in cancer cells reveals a non-linear correlation with metastatic burden

We generated an in vivo mouse model to study the impact of different *Prrx1* levels in cancer cells, by using a combination of (i) MMTV-PyMT[28] transgenic mice, which develop mammary carcinomas that progress to invasive and metastatic stages resembling human invasive breast cancer[29], (ii) a previously generated *Prrx1* cKO mouse line carrying either one (*Prrx1* +/f) or two (*Prrx1* f/f) floxed *Prrx1* alleles[19], and (iii) the Krt14-Cre tdTomato reporter line (Supplementary Fig. 2a, b). The Krt14-Cre line tags mammary gland progenitor cells from early embryonic stages[30] (Supplementary Fig. 2b). Thus, in the adult mammary gland, all epithelial cells (luminal and basal) are Tomato$^+$, regardless of Krt14 expression (Supplementary Fig. 2b, c). In PyMT tumors, virtually all cancer cells (PyMT+) are tdTomato+ (≈99% in previous studies[19]), irrespective of Krt14 expression (Supplementary Fig. 2b, d, e), while stromal cells remain negative (Supplementary Fig. 2d). Consistent with our recent finding that luminal cells acquire basal-like traits during PyMT tumor progression[19], almost 99% of K14$^+$ cells are PyMT$^+$, ruling out non-transformed basal-like cells, as the oncogene is active only in the luminal compartment (Supplementary Fig. 2b, e). Altogether, these data confirm that our model is highly sensitive and specific, enabling Prrx1 deletion in all cancer cells before tumor onset without affecting other populations.

Aware that Prrx1 has different isoforms[31], our conditional knockout (cKO) design targets exon 2, which is shared by both isoforms and encodes the critical homeobox domain. We previously showed that this strategy effectively abolishes the expression of both isoforms in embryonic fibroblasts[19]. Furthermore, heterozygous cells show similarly reduced levels of both isoforms[19]. To confirm that this is the case when combined with PyMT mice and to assess Prrx1 levels in the different allelic combinations in vivo, we analyzed both total *Prrx1* expression and that of the two isoforms (Long and Short) in FACS-sorted cancer cells from our experimental groups. We confirmed that *Prrx1* +/f cancer cells expressed approximately half the levels of Prrx1 compared to *Prrx1* +/+ cells and that both isoforms behave similarly (Fig. 2a), ruling out the possibility of differential isoform effects in our model. In addition, we found that *Prrx1* levels did not affect primary tumor burden (Fig. 2b) or the overall cell proliferation rate (Fig. 2c).

Consistent with our observations in breast cancer patients, mice bearing *Prrx1* +/f tumors exhibited the highest metastatic burden, driven by an increase in the number and the size of metastatic foci. In contrast, *Prrx1* f/f mice showed a significant reduction in metastatic burden compared to controls (Fig. 2d) in agreement with our previous findings[19]. Given that *Prrx1* promotes cancer cell invasion[19], we also assessed tumor invasiveness. In the PyMT model, luminal tumor cells progressively acquire basal-like traits, including de novo Krt14 expression, coinciding with activation of an invasive EMT program[19], and Krt14 reliably marks the invasive and disseminative population[32].

https://doi.org/10.1038/s41467-026-70242-4

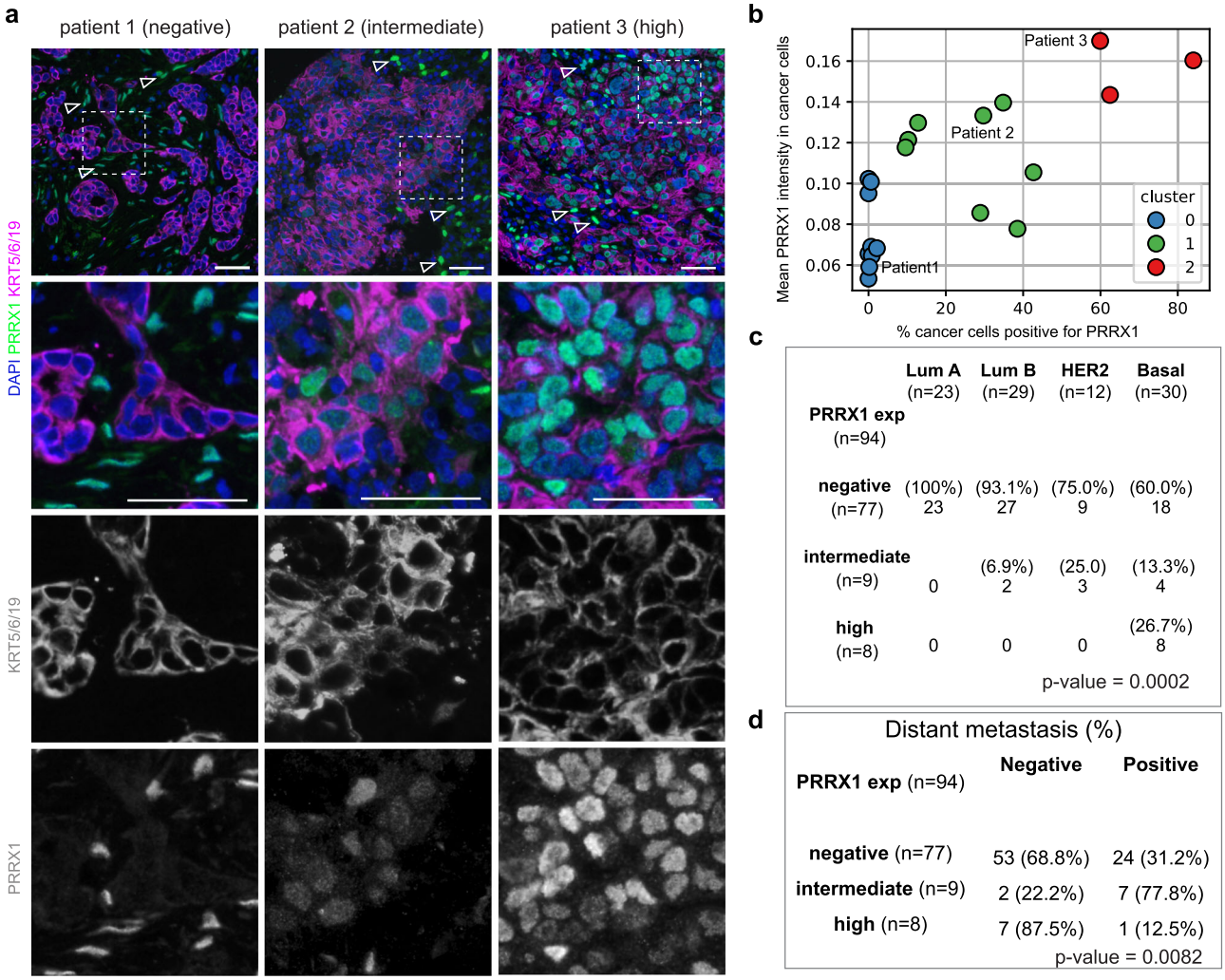


Fig. 1 | Non-linear correlation between PRRX1 levels and metastatic incidence in breast cancer patients. a Levels of Prrx1 in TMAs of breast cancer patients categorized by their expression pattern. $n = 94$ patients analyzed. Representative image from each expression pattern, see (**b**). Arrowheads indicate stromal PRRX1 expression (not in cancer cells). Scale bars: 25 μm. **b** Dot plot of patients colored by unsupervised clustering based on PRRX1 positivity and intensity in cancer cells. $n = 20$ patients. Labeled patients (1–3) represent cases shown in (**a**). **c** Correlation between subgroups of breast patients categorized by PRRX1 expression. The chi-squared test is used to determine statistically significant differences between the groups. **d** Correlation between the different groups of breast patients identified in (**b**) and incidence of distant metastasis. The chi-squared test is used to determine statistically significant differences between the groups. See Supplementary Table 4 and Source data File for additional statistical analysis. Source data are provided as a Source data file.

Therefore, we combined peripheral enrichment of Krt14 expression with the presence of disorganized laminin borders, a well-established morphological hallmark of invasive behavior, to accurately delineate invasive areas. Compatible with our data on metastatic burden, *Prrx1* +/f tumors displayed a significant increase in invasive areas with respect to control tumors, while *Prrx1* f/f tumors showed a marked reduction (Fig. 2e). This is in line with our previous data showing that Prrx1 is associated with the invasive EMT phenotype[19], and consistently, we detect Prrx1/K14 double positive cancer cells in these areas (Fig. 2f). Our data from *Prrx1* genetic breast cancer mouse models reveal a non-linear relationship between *Prrx1* levels and metastasis. Together with patient data, our findings in mouse models indicate that while a minimum threshold of *Prrx1* is required for invasion and metastasis, high levels hinder metastatic progression.

Spatial transcriptomics identifies a molecular signature of cancer cells in the invasive areas

To better characterize the invasive population, we performed single-cell spatial transcriptomics on paraffin-embedded sections from *Prrx1* +/+ primary tumors. With a custom panel of 264 genes designed to identify major cell types in PyMT tumors and capture EMT heterogeneity in cancer cells (Fig. 3a), we identified 1.794 million high-quality cells (Supplementary Fig. 3a, b), which were grouped into 18 transcriptionally distinct clusters with consistent representation across samples (Supplementary Fig. 3c). Using supervised annotation based on previously generated single-cell PyMT transcriptomics data[19], we classified cells into five major types, encompassing cancer cells and associated populations, each characterized by specific marker expression (Fig. 3b, c). As expected, cancer cells were arranged in dense islands, surrounded by tumor microenvironmental populations, some of which infiltrated the tumor (Supplementary Fig. 3d).

To analyze the invasive regions, we manually identified areas of high keratin 14 (*Krt14*) expression at tumor margins (Fig. 3d), applying a strategy similar to the quantification of invasive areas (Fig. 2e). Comparing the expression profiles of cancer cells at invasive edges (Fig. 3e) with those located elsewhere revealed that invasive cells were significantly enriched in invasion-related genes (*Tnc, Mmp2, Mmp3*) and exhibited a basal-like expression program (*Krt5, Trp63*) (Fig. 3f and Supplementary Fig. 3e), while showing lower levels of injury response

https://doi.org/10.1038/s41467-026-70242-4

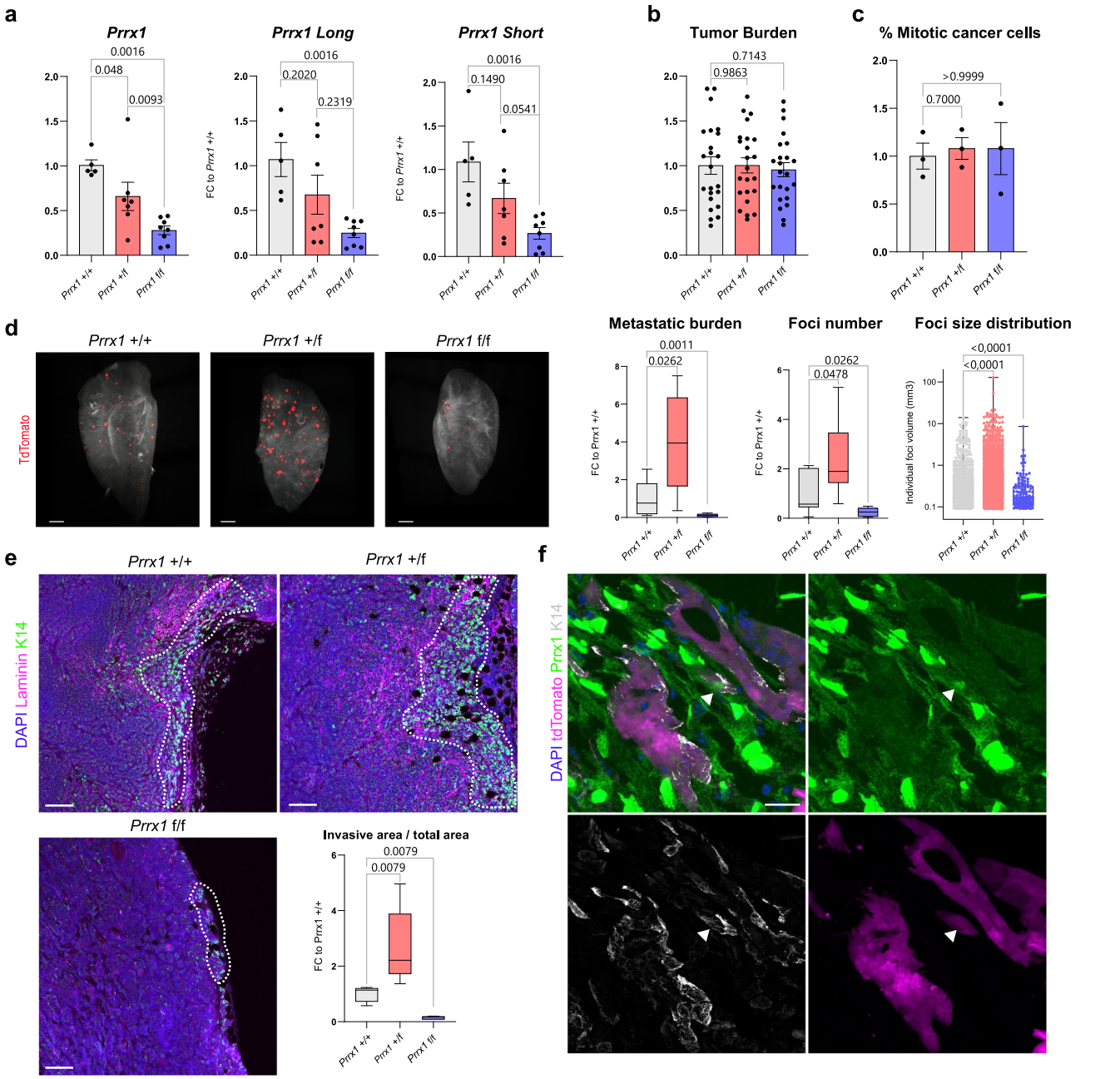


Fig. 2 | In vivo genetic modeling of different *Prrx1* levels in cancer cells reveals a non-linear correlation with metastatic burden. a RT-qPCR showing the relative transcript levels of *Prrx1* and *Prrx1* isoforms (Long and Short) in FACS-sorted cancer cells per condition. FC is represented as mean ± SEM. *Prrx1* +/+ ($n = 5$ tumors from 3 different mice); *Prrx1* +/f ($n = 7$ tumors from 4 different mice); *Prrx1* +/+ ($n = 8$ tumors from 5 different mice). Mann–Whitney two-tailed test. **b** Quantification of total tumor weight per animal between different *Prrx1* genotypes. FC is represented as mean ± SEM ($n = 23$ mice per group). Two-tailed unpaired test. **c** Mitotic ratio of cancer cells in whole PyMT tumors with different *Prrx1* genotypes. FC is represented as mean ± SEM. $n = 3$ tumors from 3 different mice. Mann–Whitney two-tailed test. **d** 3D reconstitution of whole-mounted clarified lung lobes showing metastatic cells (TdTomato), and the corresponding quantification of metastatic burden, foci number, and size ($n = 7$ mice per group). Boxplots show median, upper and lower quartiles; whiskers show upper and lower extremes. Mann–Whitney two-tailed test. Scale bars: 1000 μm. **e** Representative images and quantification of invasive areas in PyMT primary tumors. Invasive areas are defined as peripheral areas with disorganized pan-Laminin and enriched in K14-positive cancer cells (dashed lines). FC is represented as mean ± SEM; $n = 5$ mice per group. Boxplots show median, upper and lower quartiles; whiskers show upper and lower extremes. Mann–Whitney two-tailed test. Scale bars: 100 μm. **f** Image showing a Prrx1+K14+ cancer cell (white arrowhead) inside an invasive area. Representative of at least 3 different tumors. Scale bars: 50 μm. ns = $p > 0.05$, not statistically significant. Source data are provided as a Source data file.

and interferon-related gene transcripts (*Fos, Klf4, Mx1, Ifit1*) (Fig. 3f). Invasive cells exhibited an EMT transcriptional profile (Fig. 3g), associated with transcription factors such as *Snail2*, a key regulator of the basal program[33], Fibrillin-1 (*Fbn1*) and, importantly, *Prrx1* (Fig. 3h). Visual inspection of the spatial transcriptomics data confirmed that this gene program is activated in cancer cells. As several of the detected genes, including *Col1a1*, *Postn* and *Tnc*, are typically highly expressed in stromal populations, we carefully verified that these results were not biased by transcript misassignment or imperfect cell segmentation at tumor borders (Supplementary Fig. 4). Overall, our spatial transcriptomics analysis revealed a molecular signature of cancer cells located at the invasive areas, highlighting a program

https://doi.org/10.1038/s41467-026-70242-4

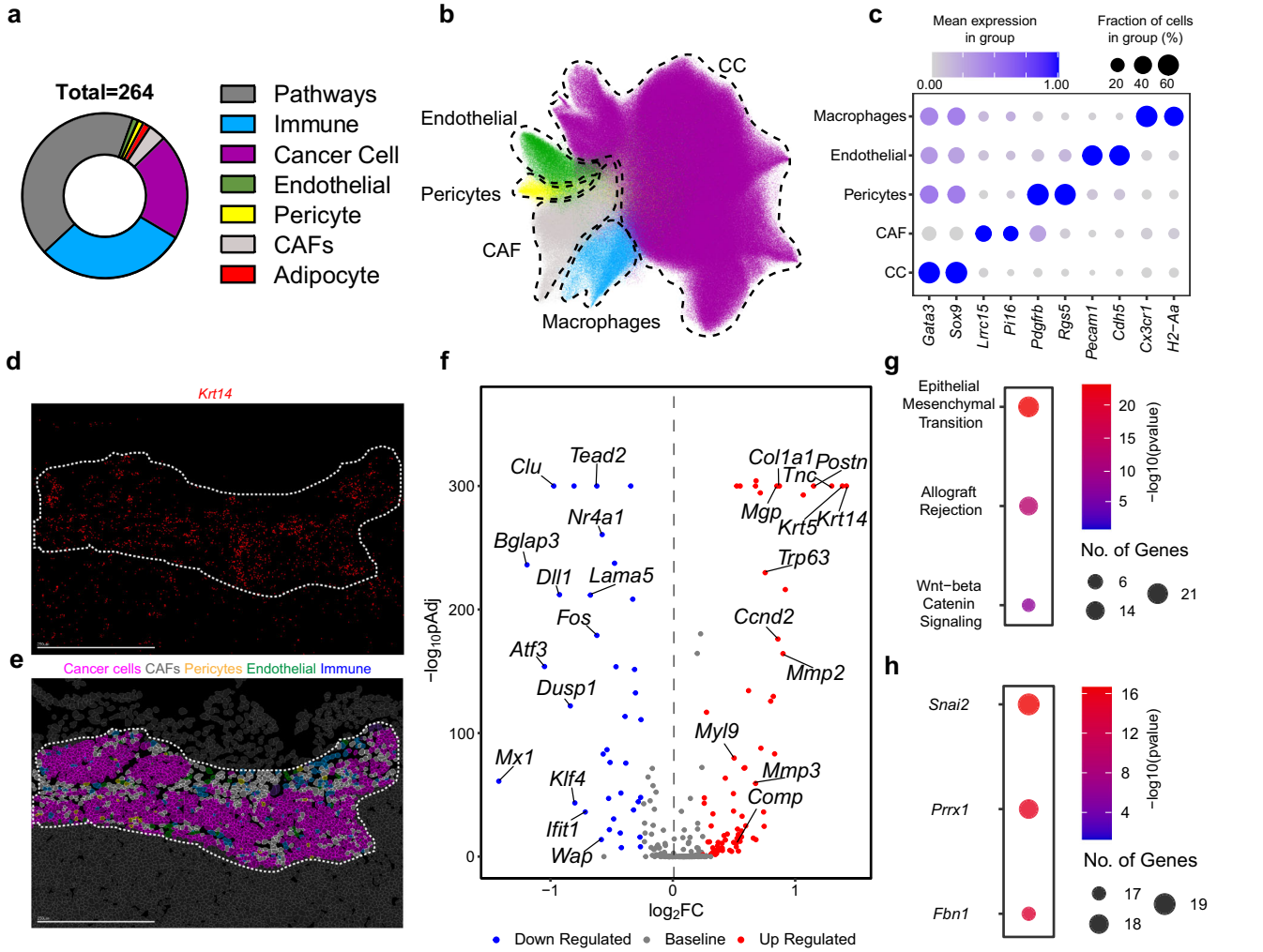


Fig. 3 | Spatial transcriptomics identifies a molecular signature of cancer cells in the invasive areas. a Distribution of the 264 Genes in the MERSCOPE Custom Panel designed to detect the main cell types and EMT activation in PyMT tumors. **b** UMAP projection of Spatial transcriptomics (ST) analysis of PyMT *Prrx1* +/+ tumors ($n = 3$), identifying different cell types. **c** Dot plot showing the expression of representative marker genes that identify different cell types, as in (**b**). **d** Example of a selected invasive edge (outlined) based on *Krt14* expression. Representative of the 13 different invasive edges (from 3 tumors from different mice) analyzed. Scale bar: 250 μm. **e** Cell types colored as in (**b**) in the selected invasive edge. Scale bar: 250 μm. **f** Volcano plot of differentially expressed genes (DEGs) in MERSCOPE-ST data for cancer cells located at the invasive edges vs the rest of the tumor ($n = 3$ tumors). DEGs with p value < 0.05 (Wilcoxon Rank Sum two-tailed test) and log2 fold change >0.25 labeled as increased (red) or log2 fold change <−0.25 labeled as decreased (blue). **g** Dot plot showing pathway enrichment (MsigDB_Hallmarks) of differentially upregulated genes ($p < 0.05$, Fisher's exact two-tailed test, log2 fold change >0.25) in cells in invasive areas. **h** Dot plot showing transcription factor association (ARCHS4) of differentially upregulated genes ($p < 0.05$, Fisher exact two-tailed test, log2 fold change >0.25) in cells in invasive areas. Source data are provided as a Source data file.

compatible with that described along the Prrx1-dependent EMT invasive trajectory in scRNA-seq datasets[19].

Intermediate levels of *Prrx1* promote an invasive, highly proliferative metastatic state

To further dissect how different *Prrx1* levels shape the transcriptional landscape of invasive cells, we performed scRNA-seq (10X Genomics) from 3 *Prrx1* +/+, 3 *Prrx1* +/f, and 4 *Prrx1* f/f tumors (see "Methods"). Unsupervised clustering identified nine major cell types and cancer cell states, each defined by bona fide lineage markers (Supplementary Fig. 5a–d). Focusing on cancer cells, we subset them and identified 28 subpopulations (Fig. 4a).

To identify the invasive population, we developed a Spatial Transcriptomics-derived Invasive Signature (STINS) using the differentially regulated genes (log2FC > 0.25 and P value < 0.05) in the cancer cells residing in the invasive areas compared to those located elsewhere (Fig. 3f; see "Methods") and applied it to the scRNA-seq data, identifying clusters 20 and 25 as enriched for invasive features (Fig. 4b). These invasive cell populations present the highest levels of EMT activation in the tumor (Fig. 4c), with downregulation of the epithelial signature and upregulation of mesenchymal genes in cluster 25, exemplified by *Epcam* and *Vim*, respectively (Supplementary Fig. 6a). This is further supported by protein expression (Supplementary Fig. 6b). Clusters 20 and 25, are enriched in a recently described pro-invasive gene signature (PINGs)[19] able to discriminate between invasive and non-invasive human breast cancer cell lines (Fig. 4d). Interestingly, these populations closely align with the final stages of the recently described invasive embryonic-like EMT trajectory in PyMT tumors (Supplementary Fig. 7a, b). Notably, *Prrx1* expression was only detected in cluster 25 (Fig. 4e).

Although we did not see differences in the distribution of the clusters (Supplementary Fig. 6c), we observed significant differences in the transcriptomic profile of cells in cluster 25 between the different genotypes. Comparative analysis of cluster 25 revealed that while the

https://doi.org/10.1038/s41467-026-70242-4

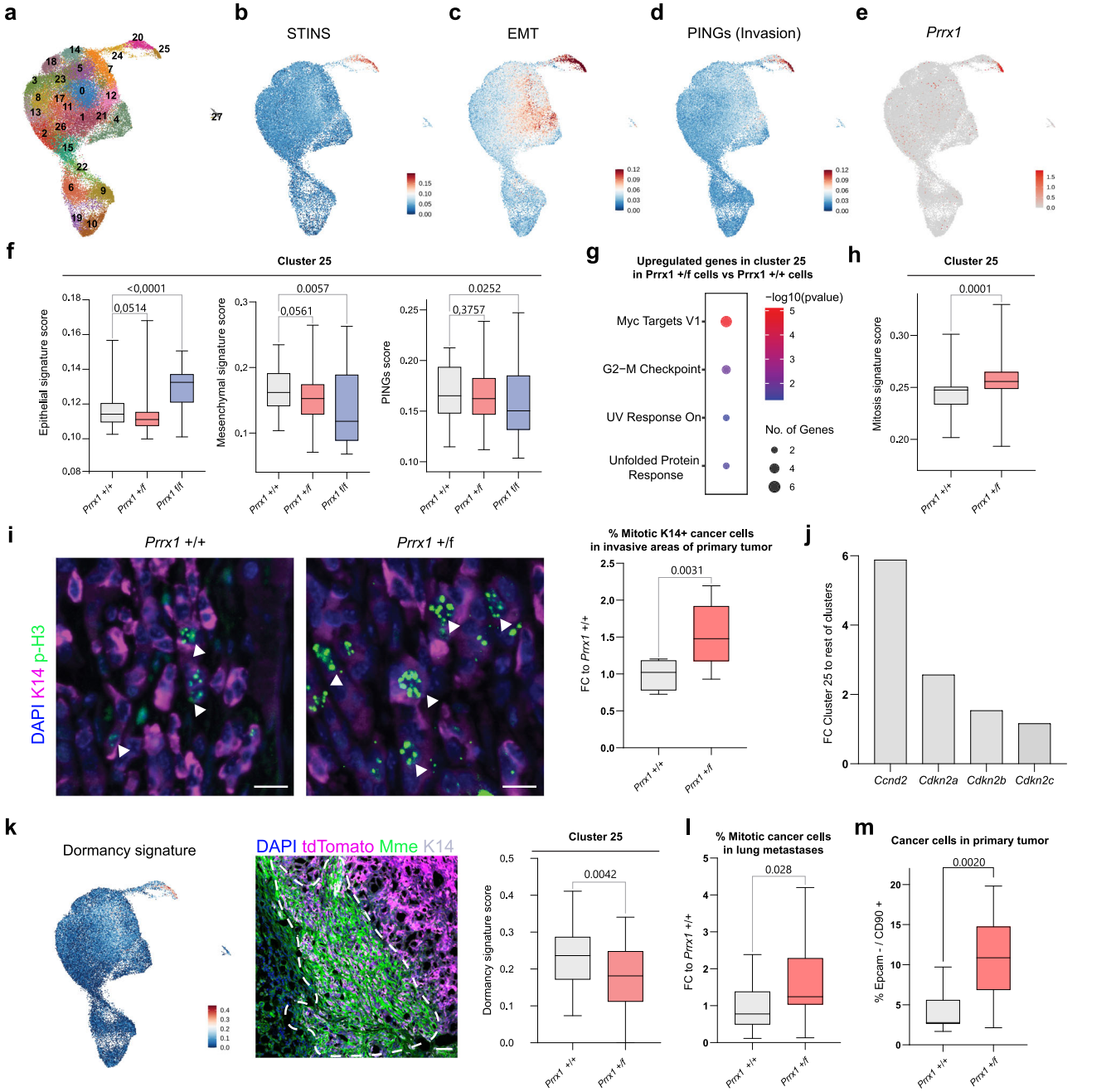


Fig. 4 | Intermediate levels of *Prrx1* promote an invasive, highly proliferative metastatic state. a UMAP projection of scRNA-seq analysis of cancer cell subset of *Prrx1* +/+ ($n = 7$); *Prrx1* +/f ($n = 3$), and *Prrx1* f/f ($n = 4$) tumors. **b** Enrichment of STINS (spatial transcriptomic-derived invasive signature) represented over UMAP. **c** UMAPs projection showing enrichment in EMT signature. **d** UMAPs projection showing the enrichment in PING (pro-invasive) signature[19]. **e** Normalized expression of *Prrx1* represented over UMAP. **f** Score of Epithelial, Mesenchymal and PING signatures of cells in cluster 25 of *Prrx1* +/+ ($n = 34$); *Prrx1* +/f ($n = 78$), and *Prrx1* f/f ($n = 66$) cells from *Prrx1* +/+ ($n = 3$); *Prrx1* +/f ($n = 3$), and *Prrx1* f/f ($n = 4$) tumors. Boxplots show median, upper and lower quartiles, and whiskers show upper and lower extremes. Mann–Whitney two-tailed test. **g** Dot plot showing pathway enrichment (MsigDB_Hallmarks) analysis of differentially upregulated genes ($p < 0.05$, Fisher exact two-tailed test) determined by MUSCAT of *Prrx1* +/f vs *Prrx1* +/+ cells in cluster 25. **h** Score of cell cycle signatures (Cell_Cycle_Mitotic_R-HSA-69278) on cluster 25 of *Prrx1* +/+ ($n = 34$) and *Prrx1* +/f ($n = 78$) cells from *Prrx1* +/+ ($n = 3$) and *Prrx1* +/f ($n = 3$) tumors. Boxplots show median, upper and lower quartiles; whiskers show upper and lower extremes. Mann–Whitney two-tailed test. **i** Representative images and quantification of the mitotic rate of K14+ cancer cells at the invasive edges of primary tumors. $n = 9$ edges of *Prrx1* +/+ tumors from 3 mice; $n = 11$ edges of *Prrx1* +/f tumors from 3 mice. Boxplots show median, upper and lower quartiles, and whiskers show upper and lower extremes. Mann–Whitney two-tailed test. Scale bar: 50 μm. **j** Enrichment in the expression of cell cycle regulators in cluster 25 (scRNA-seq) relative to that in all the other clusters. **k** Dormancy signature represented over the UMAP. Representative image showing Mme (part of the dormancy signature) expression in tumors, showing specificity for K14+ invasive areas. Scale bars: 50 μm. Score of Dormancy signature of cells in cluster 25 of *Prrx1* +/+ ($n = 34$) and *Prrx1* +/f ($n = 78$) cells from *Prrx1* +/+ ($n = 3$) and *Prrx1* +/f ($n = 3$) tumors. Boxplots show median, upper and lower quartiles; whiskers show upper and lower extremes. Mann–Whitney two-tailed test. **l** Mitotic rate of cancer cells in lung metastases. *Prrx1* +/+ $n = 24$ metastases from 7 mice; *Prrx1* +/f $n = 19$ metastases from 6 mice. Boxplots show median, upper and lower quartiles; whiskers show upper and lower extremes. Two-tailed unpaired test. Scale bar: 25 μm. **m** Quantification by FACs of CD90 status in Epcam$^-$ cancer cells from *Prrx1* +/+ ($n = 9$ tumors from 3 mice) and *Prrx1* +/f ($n = 14$ tumors from 4 mice) tumors. Boxplots show median, upper and lower quartiles; whiskers show upper and lower extremes. Mann–Whitney two-tailed test. ns = $p > 0.05$, not statistically significant. Source data are provided as a Source data file.

complete loss of *Prrx1* (*Prrx1* f/f) markedly reduced mesenchymal features and invasive signatures, *Prrx1* +/f cells retained an EMT phenotype similar to the *Prrx1* +/+ cells despite their lower level of *Prrx1* expression (Fig. 4f). In addition, these findings indicate that intermediate *Prrx1* levels are sufficient to sustain an EMT-driven invasive program, compatible with a partial and invasive EMT phenotype, previously described as highly metastatic[15,16,18].

To better understand the increased metastatic potential of *Prrx1* +/f tumors, we compared the expression profiles of cells in cluster 25 in *Prrx1* +/+ and *Prrx1* +/f conditions. Pathway enrichment analysis revealed that genes upregulated in *Prrx1* +/f cells were primarily associated with proliferation and cell cycle regulation (Fig. 4g). This was further confirmed using cell cycle scoring (Reactome), which indicated that *Prrx1* +/f invasive cells are significantly more proliferative than their *Prrx1* +/+ counterparts (Fig. 4h). Immunofluorescence quantification of cycling cancer cells at the invasive edge in *Prrx1* +/+ and *Prrx1* +/f tumors validated this observation (Fig. 4i), compatible with the enlarged invasive areas observed in *Prrx1* +/f tumors (Fig. 2e). Importantly, this effect on proliferation is restricted to the invasive cell population. As expected from an effect dependent on Prrx1 dose, we did not observe significant changes in primary tumor size and global proliferation (Fig. 2b, c), as over 95% of tumor cells do not express Prrx1 (Fig. 4e)

To identify potential regulators of changes in proliferation in cluster 25, we examined the expression of cell cycle-related genes in invasive cells and found a notable enrichment of Cyclin D2 and CDKN2 family members when comparing cells in cluster 25 with the rest (Fig. 4j). Given that CDKN2 proteins are key inhibitors of the cell cycle[34], and cell cycle arrest is one of the main characteristics of dormancy, we investigated whether the *Prrx1*-positive invasive cell population also expressed dormancy-associated genes. Using a breast cancer dormancy signature recently derived by comparing cycling and non-cycling metastatic PyMT cells, able to differentiate between the dormant and non-dormant version of the same cells (D2.A1 vs D2.OR cell lines[35]), we found that our invasive clusters had the highest dormancy scores in the tumor, cluster 25 in particular (Fig. 4k). Immunofluorescence analysis of Mme expression, a marker of the dormancy program, confirmed high expression in invasive areas (Fig. 4k), indicating that dormancy and invasion are coregulated in tumors. Notably, *Prrx1* +/f cells in cluster 25 exhibited a significantly lower dormancy score than their *Prrx1* +/+ counterparts (Fig. 4k), suggesting that intermediate *Prrx1* levels allow invasive cells to avoid the activation of dormancy. To assess whether these differences in proliferation and dormancy at the invasive edge could be translated to the metastatic site, we quantified the proportion of proliferating cancer cells in lung metastases, a common approach to evaluate dormancy at the metastatic site[36,37]. Metastases derived from *Prrx1* +/f tumors displayed a significantly higher fraction of cycling cells compared to those from *Prrx1* +/+ tumors (Fig. 4l), consistent with the previously observed increased size of *Prrx1* +/f metastatic foci (Fig. 2d). Finally, we aimed to determine whether this state was also associated with enhanced tumor-initiating capacity using CD90, a well-characterized marker of metastatic-initiating cells in the PyMT model[38]. FACS analysis confirmed that the Epcam$^-$ cancer cell population—and cluster 25 shows the lowest *Epcam* expression in the tumor (Supplementary Fig. 6a) was enriched in CD90$^+$ cells in *Prrx1 +/f* tumors compared to the same population in control *Prrx1* +/+ tumors, suggesting that intermediate *Prrx1* levels promote metastatic-initiating capabilities (Fig. 4m). Interestingly, this finding aligns with previous observations in cultured human cell lines, where *PRRX1* downregulation enhanced sphere formation increasing the proportion of CD44$^+$/CD24$^-$ cells[20]. Overall, these findings suggest that while high *Prrx1* expression maintains invasive cells in a dormant, low-proliferative state, intermediate levels sustain invasive potential while allowing proliferation, facilitating metastatic colonization. This mechanism could explain the higher metastatic incidence observed in patients bearing breast tumors with intermediate PRRX1 levels.

Chromatin accessibility analysis establishes *Prrx1* as a direct regulator of proliferation and dormancy in the invasive population

To further investigate how *Prrx1* could regulate cell cycle progression and dormancy, we profiled chromatin accessibility after snATAC-seq and analyzed chromatin dynamics in *Prrx1* control tumors (Fig. 5 and Supplementary Figs. 8 and 9). Unsupervised clustering of open chromatin data identified six main clusters (Supplementary Fig. 9a), with a consistent cell distribution across samples (Supplementary Fig. 9b). We used gene activity of bona fide cell type markers to annotate clusters into major cell types (Supplementary Fig. 9c, d). As for the scRNA-seq analysis, we focused on the cancer cells and identified 16 distinct clusters (Fig. 5a). To identify invasive cells, we imputed gene expression in sATAC-seq cells using chromatin accessibility data and scRNA-seq as a reference (see "Methods") and applied STINS. As observed in the scRNA-seq data, two clusters, 14 and 15, emerged as the invasive populations (Fig. 5b), with cluster 15 exhibiting higher *Prrx1* imputed expression (Fig. 5c), mirroring cluster 25 in the scRNA-seq analysis. As observed in the scRNA-seq dataset, we also observed a correlation between chromatin states and our recently described EMT trajectories, indicating that distinct transcriptional EMT states are supported by specific chromatin profiles (Supplementary Fig. 10a, b).

Given the nature of Prrx1 as a transcription factor, we hypothesized that it could be a master regulator of the invasive and metastatic cells. Transcription factor binding motif analysis showed that Prrx1 binding sites are more accessible in cluster 15 compared to the rest of the clusters (Fig. 5d). Further analysis revealed that several accessible regions in cluster 15, located near the promoters of genes encoding cell-cycle regulators (cyclins and cyclin inhibitors) and dormancy-related genes[34] (*Gas6*, *Mme*), respectively, enriched in the invasive cell population in the scRNA-seq dataset (Fig. 4j) and part of the dormancy signature (Fig. 4k), are positive for Prrx1 binding motifs as assessed with the "Find Individual Motif Occurrences" tool (FIMO; Fig. 5e, f).

These results indicate that Prrx1 is not only specifically expressed in the invasive population but also that it plays an important regulatory role, as it can directly bind promoters of genes associated with cell-cycle control and dormancy, providing mechanistic evidence for the differences in metastatic potential observed at different Prrx1 levels.

Engineered cell line models reveal dose-dependent Prrx1 regulation of invasion, proliferation, and dormancy

To examine whether this regulatory pattern was conserved in human, we analyzed whole-genome transcriptomes of human breast cancer cell lines[39] ordered and grouped by EMT status[19] (Fig. 6a). Firstly, we compared PRRX1 expression across EMT groups and confirmed that the mesenchymal group shows the highest levels (Supplementary Fig. 11a), consistent with PRRX1 being enriched in mesenchymal cells and expressed in PyMT tumors in cells with the highest EMT score[19]. Hybrid E/M groups displayed intermediate expression when compared to epithelial and mesenchymal states (Supplementary Fig. 11a).

When scoring the different gene signatures in this dataset, a clear correlation emerged between *PRRX1* expression, mesenchymal character, invasion (PING) and dormancy signatures (Fig. 6a). To assess its predicted regulatory role, we selected BT549 cells, as they express high levels of *PRRX1* and of dormancy and cell cycle inhibitor genes (Fig. 6a), the closest to cells in cluster 25 from the scRNA-seq data in PyMT tumors. In addition, we had previously shown that *PRRX1* downregulation is necessary for BT549 cells to metastasize, while it reduced the invasive capacity of the BT-549 cells[20], making them an ideal model for assessing the impact of different *PRRX1* downregulation on cell cycle regulators and dormancy-related genes.

https://doi.org/10.1038/s41467-026-70242-4

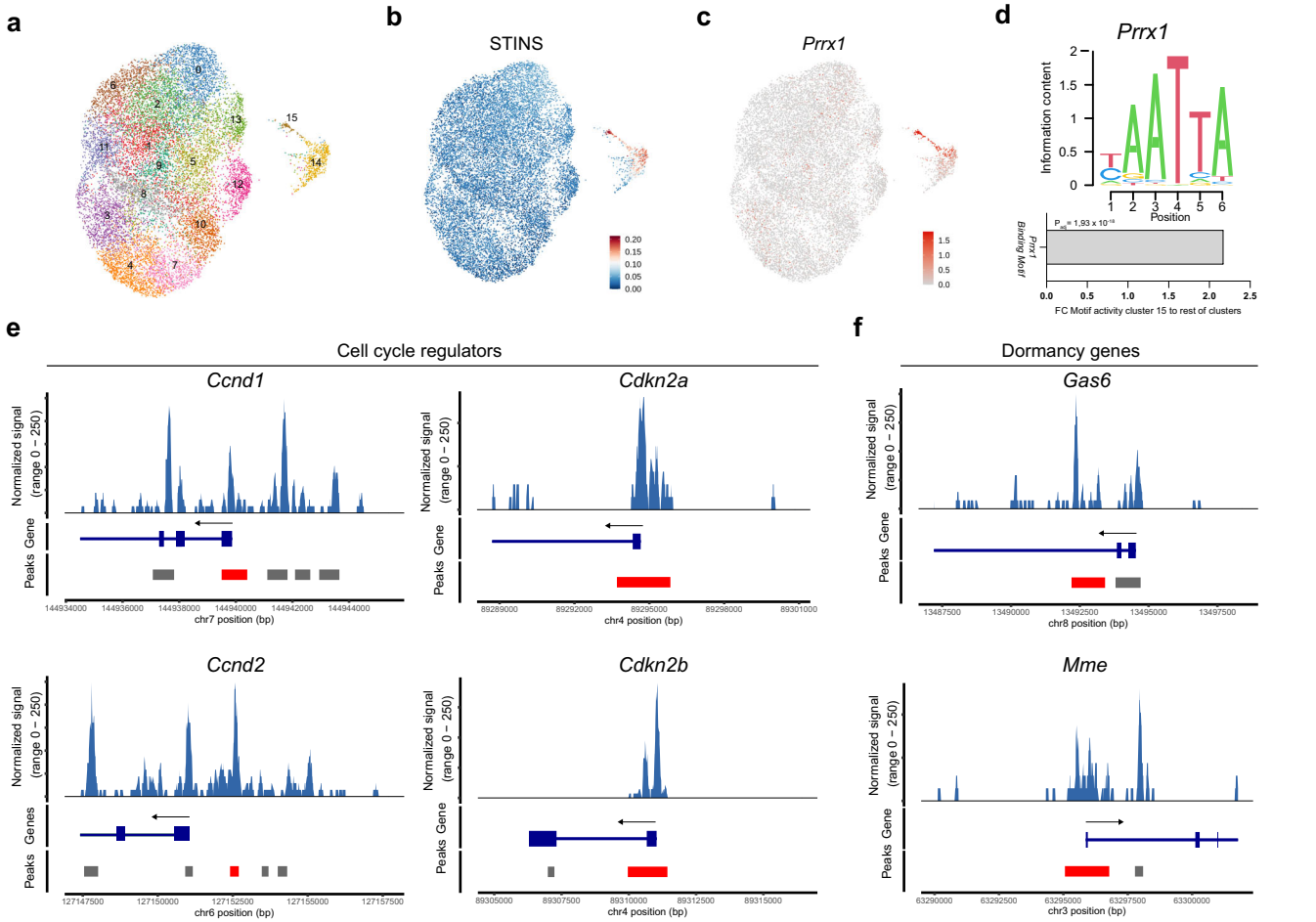
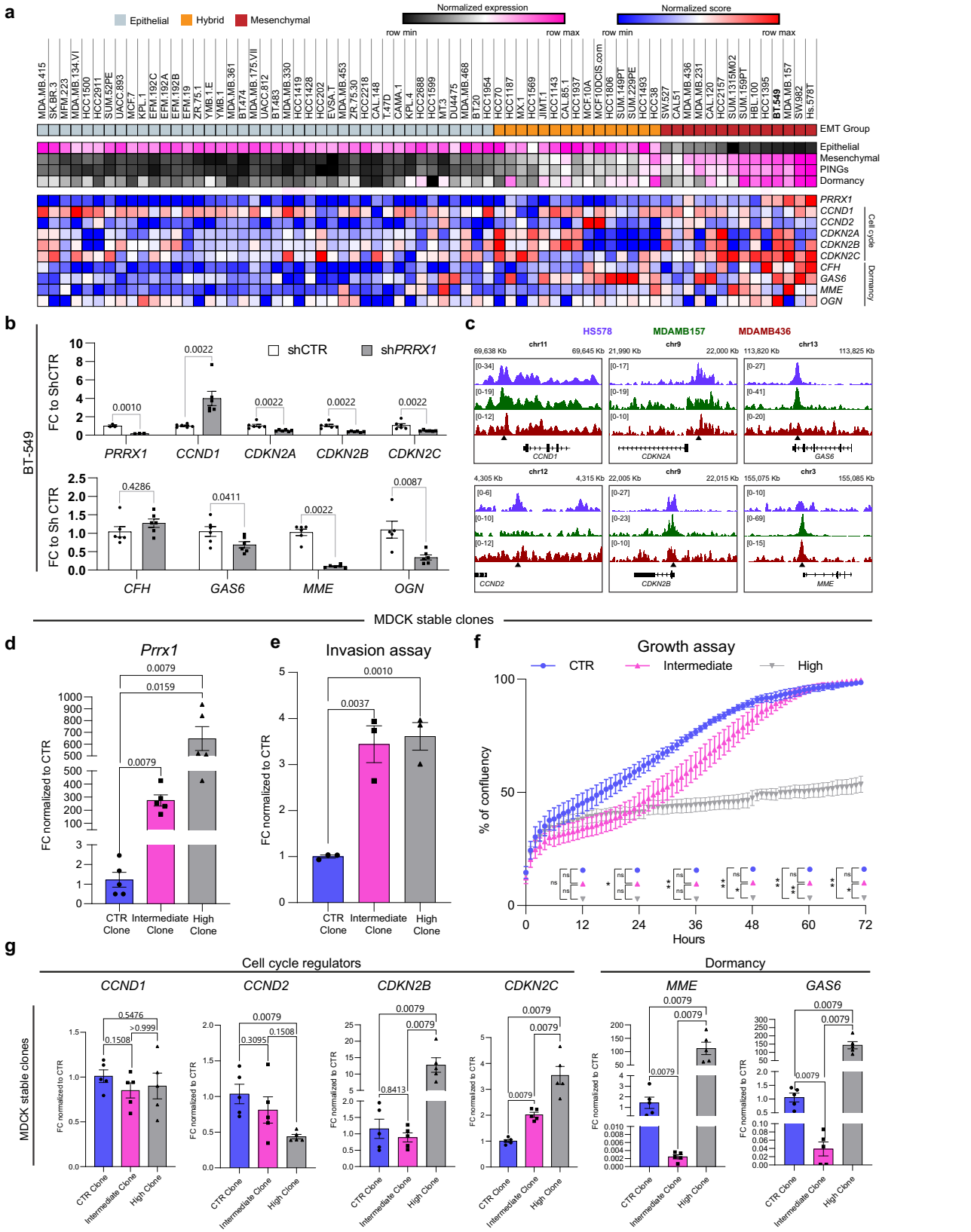


Fig. 5 | Single cell chromatin accessibility analysis establishes *Prrx1* as a direct regulator of proliferation and dormancy in the invasive population. a UMAP of cancer cells from *Prrx1* +/+ tumors ($n = 4$) as assessed by sATAC-seq analysis. **b** Enrichment of STINS (spatial transcriptomic-derived invasive signature) based on imputed expression (see "Methods") represented over the UMAP. **c** Normalized imputed expression of Prrx1 over the UMAP. **d** Prrx1 binding motif and enrichment of motif activity in cluster 15 compared to the rest of the clusters (Wilcoxon Rank Sum two-tailed test). **e** Coverage plots of identified peaks in cells from cluster 15 linked to genes encoding regulators of cell proliferation: cyclins (*Ccnd1* and *2*) and cyclin Inhibitors (*Cdkn2a* and *2c*). Peaks with identified binding sites for Prrx1 are indicated in red. **f** Similar analysis as that in (**e**) but with dormancy-associated genes. Source data are provided as a Source data file.

Downregulation of *PRRX1* by RNA interference (BT549 sh*PRRX1* cells) led to an increase in the expression of the *CCND1* cyclin and a reduction in that of cyclin inhibitors of the *CDKN2* family and dormancy-associated genes (*GAS6, MME, OGN*) (Fig. 6b). In addition, analysis of published data on PRRX1 downregulation in the HS578T human breast cancer cell line[40] revealed a clear upregulation of *CCND2*, another Cyclin D family member, while changes in *CCND1* were less evident, along with downregulation of *MME*, a dormancy-associated gene (Supplementary Fig. 11b). Finally, the analysis of ChIP-Seq dataset in three different human breast cancer cell lines with different *PRRX1* levels[40], including HS578T, shows that, at least one of them, PRRX1 can directly bind the promoters of cell cycle regulators, dormancy and invasion genes in human breast cancer cell lines (Fig. 6c and Supplementary 11c). These findings confirm that *PRRX1* is an activator of cell cycle inhibitors and dormancy-associated genes in human breast cancer and provide an explanation for the requirement of PRRX1 downregulation for metastatic outgrowth.

To mechanistically model graded Prrx1 expression, we engineered cell-line variants expressing low (CTR), intermediate, or high levels of Prrx1 (Fig. 6d). The difference between intermediate and high expression was approximately twofold, closely matching the allelic differences observed in our mouse model. We selected MDCK-NBL2 epithelial cells as the parental line because of their pronounced phenotypic plasticity, well-characterized EMT response to TGF-β[19,41–44], and their ability to recapitulate the sequential activation of EMT transcription factors observed in neural crest cells and in the PyMT model, including the late induction of Prrx1[19]. Importantly, the MDCK system allows us to study the effects of increasing Prrx1 dosage on cell-state transitions and to directly assess its relationship with invasion and proliferation. Both intermediate- and high-Prrx1-expressing clones displayed similar and significant invasiveness compared to CTR cells (Fig. 6e). In contrast, proliferation assays revealed a marked reduction in growth only in Prrx1-high clones (Fig. 6f). Gene expression analysis mirrored these functional outcomes. Prrx1-high clones exhibited reduced expression of *CCND1* and increased levels of *CDKN2B/C*, whereas intermediate clones remained closer to CTR cells (Fig. 6g). Consistent with this, dormancy-associated genes (*GAS6*, *MME*) were strongly upregulated in Prrx1-high clones (~100-fold) but downregulated in intermediate clones (Fig. 6g). Collectively, these data indicate that high Prrx1 levels promote invasion while suppressing proliferation and inducing dormancy, whereas intermediate Prrx1 expression enhances invasion with minimal impact on proliferation, thereby maximizing metastatic potential. These findings recapitulate our in vivo observations and support a dose-dependent, hormetic model of Prrx1 function regarding the regulation of metastatic potential.

https://doi.org/10.1038/s41467-026-70242-4



Patient stratification by combining invasion and proliferation signatures reveals a high-risk breast cancer subgroup with intermediate Prrx1 levels

We next wanted to assess whether the relationship between *PRRX1* levels and invasion, proliferation, and dormancy observed in the mouse tumors and human cell lines was conserved in human breast cancer. We analyzed the human breast cancer METABRIC database[24,25] and found that *PRRX1* expression positively correlates with signatures associated with invasion (PINGs) and dormancy, and negatively correlated with proliferation (Fig. 7a). Notably, this relationship extends to genes identified as putative PRRX1 targets in our previous analyses, including *MMP2,*

Fig. 6 | Engineered cell line models reveal dose-dependent Prrx1 regulation of invasion, proliferation and dormancy. a Expression analysis of signature and genes of interest in 71 breast cancer (BC) cell lines ordered according to their epithelial to mesenchymal score[19]. **b** RT-PCR of *PRRX1 and* cell cycle regulators levels (left) and dormancy-related genes (right) in shCTR vs *shPRRX1 BT549* cells. FC is represented as mean ± SEM (*PRRX1* $n = 3$ biological replicates, *OGN shCTR* $n = 5$ biological replicates, rest = 6 biological replicates, 3 technical replicates each). Biological replicate: different passage / well and extracted RNA. Technical replicate: same RNA analyzed 3 times. Mann–Whitney two-tailed test. **c** Peaks representing PRRX1 binding at the genomic loci of cell cycle regulators and dormancy genes in 3 human breast cancer cell lines: HS578 in blue, MDAMB157 in green, MDAMB436 in red. **d** RT-qPCR showing the relative *Prrx1* transcript levels in transfected and stable MCDK-NBL2 clones. FC is represented as mean ± SEM ($n = 5$, biological replicates, 3 replicates each). Biological replicate: different passage and extracted RNA. Technical replicate: same RNA analyzed 3 times. Mann–Whitney two-tailed test. **e** Quantification of invasive cells in Transwell assays in the *Prrx1*-expressing clones. FC is represented as mean ± SEM ($n = 3$ independent biological replicates). Biological replicate: different passage and independent invasion assay. Unpaired two-tailed *t*-test. **f** Growth analysis of the different *Prrx1*-expressing clones. Each time-point is represented as mean ± SEM ($n = 3$ biological replicates and 3 technical replicates each). Biological replicate: different passage, and independent imaging. Technical replicate: same condition plated in the same plate and imaged at the same time. Two-way ANOVA test, precise *p* values in Source data File. **g** RT-PCR analysis of transcripts for cell cycle regulators (left) and dormancy-related genes (right) in the *Prrx1*-expressing clones. FC is represented as mean ± SEM ($n = 5$ independent biological replicates, 3 technical replicates each). Mann–Whitney two-tailed test. ns = $p > 0.05$, not statistically significant. Source data are provided as a Source data file.

ADAMTS2, FBN1, COL1A1, CCND1, CDKN2B/C, CFH, GAS6, OGN, MME (Supplementary Fig. 12).

To evaluate whether the interplay between invasion and proliferation has prognostic value, we stratified METABRIC patients[5,24,25] into four groups (see "Methods") based on their scores for the invasion (PING)[19] and proliferation (Cell_Cycle_Mitotic_R-HSA-69278) signatures, both of which were found to be dependent on *PRRX1* levels in the invasive population in our genetic mouse models (Fig. 4f, h).

Interestingly, the impact of high invasion on survival was contingent on proliferation status; patients with highly invasive but low-proliferative tumors showed better prognosis than the rest of the groups (Fig. 7b and Supplementary Table 4). More importantly, this classification revealed that patients with tumors scoring high for both invasion and proliferation exhibit significantly worse overall survival (OS), particularly during the first months after diagnosis (Fig. 7b and Supplementary Table 4), compatible with our results modeling different *Prrx1* levels in the mouse tumors and with our proposal that metastatic competence arises from the coexistence of both programs, invasion and proliferation. Importantly, even when stratified by treatment, a very similar pattern emerged across all conditions (Supplementary Fig. 13).

Given the established link between invasion and proliferation, we next investigated whether our patient stratification reflected differences in dormancy scores and found that, as expected, patients from the group with the poorest survival, highly invasive and highly proliferative tumors showed a markedly lower dormancy score (Fig. 7c). Interestingly, this group of aggressive tumors show lower *PRRX1* expression levels than the invasive, low-proliferative group (Fig. 7d). These data fit with the non-linear metastatic burden observed in TMAs from patients with different PRRX1 expression levels (Fig. 1). Taken together, these findings indicate that human breast cancer patients with tumors showing intermediate levels of PRRX1 in cancer cells exhibit a unique combination of invasive and proliferative traits, which defines a particularly aggressive subset of tumors. Moreover, our stratification approach could serve as a valuable tool to identify high-risk patients who may benefit from therapeutic strategies targeting both invasion and proliferation.

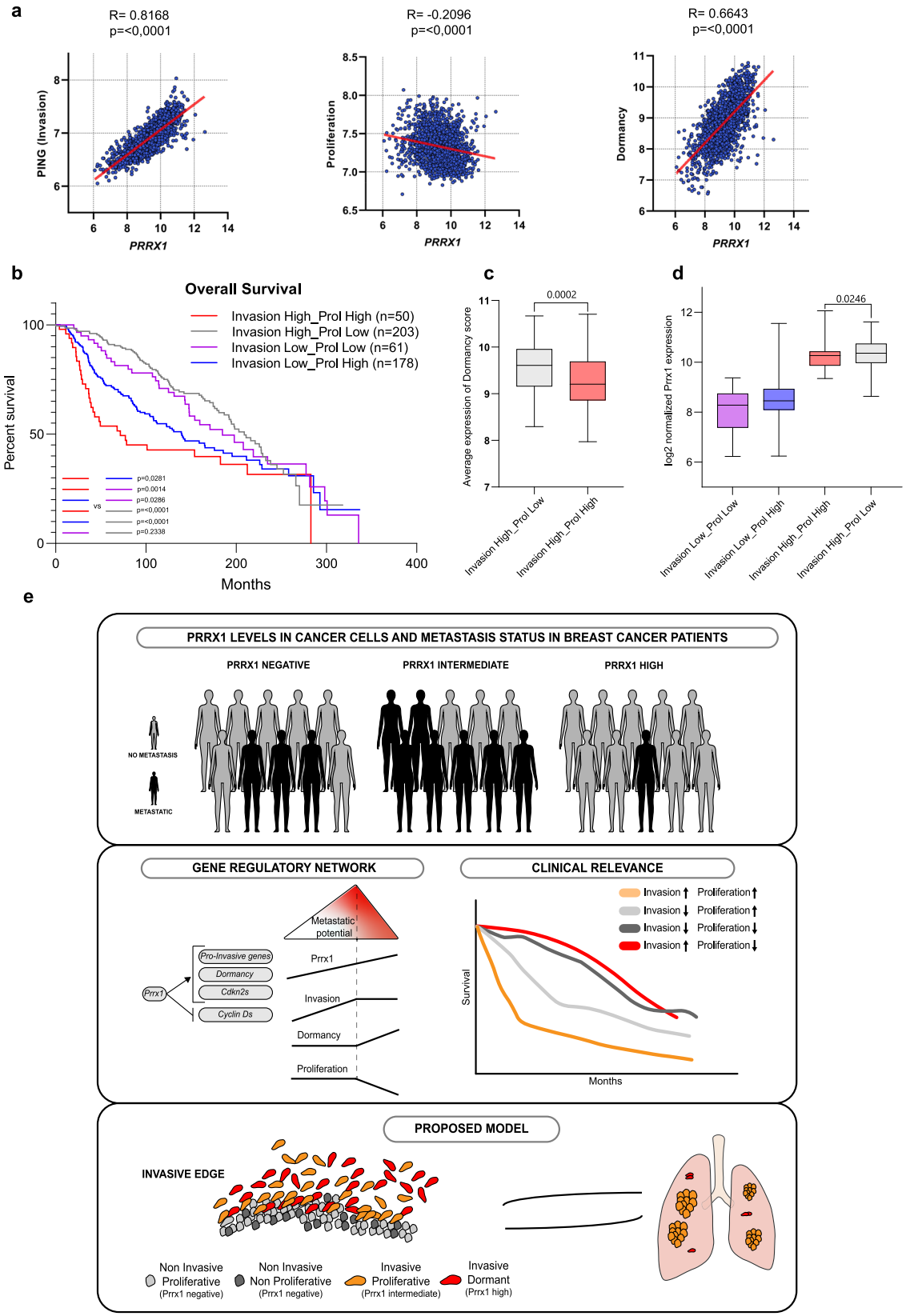
Discussion

Understanding and fighting metastasis remains one of the most significant clinical challenges in cancer. Identifying key metastatic determinants is difficult due to the complexity and multi-step nature of the metastatic cascade[6]. In this study, we integrate human breast cancer data, genetic mouse models, and multi-omics analysis to identify a particular cell state in the primary tumor that predicts high metastatic outgrowth determined by Prrx1 levels, providing insights into how the EMT status of invasive cells influences their metastatic potential.

We have previously shown that EMT is an important driver of cell plasticity in breast cancer, establishing an additional level of tumor heterogeneity by engaging cells in two distinct EMT trajectories: an embryonic-like EMT associated with invasion and an adult-like EMT associated with inflammation and immune regulation[19]. Here, we identify the EMT-TF Prrx1 as a master regulator of the metastatic potential established at the primary tumor. Prrx1 levels directly correlate with both invasion and dormancy, revealing that cells with high expression cannot translate their highly invasive and disseminative properties into metastatic outgrowth. These findings explain why intermediate Prrx1 levels (and intermediate EMT), simultaneously sustaining invasion and proliferation, hold the highest metastatic potential, and provide mechanistic insight into (i) the inverse correlation that, like with Snail1, exists between Prrx1 expression and cell cycle progression[20,21,45] (ii) the high metastatic potential of hybrid E/M cells[15–18,46], and (iii) the recent clonal analyses in human cancer linking metastasis to highly proliferating clones in the primary tumor[47,48]. This unexpected finding, given that disseminating cells undergoing EMT have been associated with a less proliferative phenotype, can be interpreted as proposing the existence of rare clones within the EMT populations that are both highly proliferative and highly invasive, compatible with intermediate Prrx1 levels serving as the main source of metastases.

There are examples in which cells respond to different levels of a signal/factor by activating different programs in development or in cancer, such as the response to Sox2 levels driving the developmental switch between stemness and differentiation[49] or how different levels of Twist regulate stemness or cancer cell invasiveness[50] but what we describe here is a non-linear correlation between expression levels and the regulation of a single trait, metastatic potential. This biphasic dose–response is known as hormetic and has been studied in the fields of toxicology and cellular stress[22,51], where several proteins, including the Nrf-2 and NF-κB transcription factors, have been identified as hormetic[52]. Hormesis can explain why different Nrf-2 levels can have opposing effects on lung cancer initiation and progression[53]. Altogether, the hormetic behavior of cancer regulators underscores the importance of considering expression levels of cancer drivers and biomarkers when designing therapeutic strategies. An example with clear clinical impact of intermediate levels is the case of the HER2-low breast cancer subtype, which has dramatically changed the therapeutic options of almost 55% of breast cancer patients[54]. Here, we identify intermediate PRRX1 expression as a prospective marker of metastatic potential independent of treatment modality. However, large, prospective, multicenter studies specifically designed to assess treatment response and long-term survival are needed to establish PRRX1 as a clinical decision-making or predictive biomarker. In addition, investigating its onset of expression and its value in stratifying less aggressive breast cancer tumors (grades 1 and 2) is a promising future direction.

https://doi.org/10.1038/s41467-026-70242-4



Most studies rely on experimental metastasis models using cell lines, which cannot reproduce the intratumor phenotypic heterogeneity observed in vivo and its impact on different steps along the metastatic cascade. Here, we show that heterogeneity at the invasive edge of the primary tumor, driven by different Prrx1 levels, is sufficient to determine metastatic behavior, and that cancer cells leave the primary tumor with intrinsic information that primes their metastatic potential. Cells expressing intermediate levels of *Prrx1* can maintain proliferation while invading and disseminating, enabling rapid metastatic outgrowth. On the other hand, cells expressing high *Prrx1* levels, albeit more invasive, would enter dormancy and require additional signals from the metastatic niche to attenuate their EMT

Fig. 7 | Patient stratification by combining invasion and proliferation signatures reveals a high-risk breast cancer subgroup with intermediate *PRRX1* levels. a Pearson correlation analysis between *PRRX1* expression and PING, Proliferation and Dormancy signatures on METABRIC human breast cancer data ($n = 1980$). The red line represents the regression line. Two-tailed Pearson correlation test. **b** Kaplan–Meier overall survival (OS) analysis of METABRICT breast cancer patients stratified by high and low expression of invasion (PINGs) and proliferation (Cell_Cycle_Mitotic_R-HSA-69278) signatures ($n = 492$) (see "Methods"). Gehan-Breslow-Wilcoxon two-tailed test. See Supplementary Table 4 and Source data file for additional statistical analysis. **c** Dormancy score in the same patient groups defined in (**b**). Boxplots show median, upper and lower quartiles; whiskers show upper and lower extremes. Mann–Whitney test. **d** Comparison of *PRRX1* levels in the same patient groups described in (**b**). Boxplots show median, upper and lower quartiles; whiskers show upper and lower extremes. Mann–Whitney two-tailed test. **e** Prrx1 levels discriminate between proliferative and dormant invasive states. Patients with breast tumors bearing intermediate levels of Prrx1 have a higher metastatic incidence than those with negative or high levels. Prrx1 heterogeneity in the primary tumor defines the metastatic potential based on cell invasion and proliferation capabilities. Mechanistically, Prrx1 activates a pro-invasive gene program while promoting cell cycle arrest and dormancy through the activation of the cell cycle inhibitors of the CDKN2 family members and the repression of Cyclin Ds. The combination of invasion plus proliferation/dormancy index leads to a hormetic response of Prrx1 levels with respect to metastatic potential. The combination of Prrx1-regulated invasion and proliferation programs allows breast cancer patient stratification, identifying a subgroup with the poorest prognosis. ns = $p > 0.05$, not statistically significant. Source data are provided as a Source data file.

(mesenchymal) state and revert to a more epithelial phenotype to resume growth[36,37]. This heterogeneity can explain previous observations of differential sensitivity to pro-dormancy signals in the metastatic niche[11,36], as the dormancy program can already be activated in the primary tumor. This explains why some cells can rapidly resume growth while others enter dormancy.

The coexistence of invasion and proliferation as a driver of metastasis enables combinatorial stratification strategies. Using this approach in breast primary tumors, we identified a subgroup of patients with poor prognosis, which may not be unique to breast cancer. In addition, assessment of the proliferative status of circulating tumor cells (CTCs) could further refine patient classification, as recent data show that patients with high expression of the proliferation marker Ki-67 in CTCs have significantly reduced progression-free survival[55]. The relationship we have found between Prrx1 levels and dormancy can also be considered in strategies aiming at maintaining dormancy[56] to prevent metastasis outgrowth, particularly in patients subjected to primary tumor resection.

In summary, our study reveals that metastatic competence is already established in the primary tumor. In particular, we find that different levels of Prrx1 create cancer cell heterogeneity at the invasive edge, determining metastatic competence in a non-linear manner by regulating invasion, proliferation, dormancy and tumor-initiating capacity (Fig. 7e). This provides an explanation for how some cancer cells enter dormancy while others rapidly proliferate in the same microenvironment and integrates previous findings on the role of Prrx1 all throughout the metastatic cascade. It is a key promoter of invasion but can act as an inhibitor of metastatic colonization[20]. Our data also provide mechanistic insight into the low metastatic potential of mesenchymal cells by showing that the EMT-TF Prrx1, while required for invasion, directly regulates cell-cycle inhibitors and the dormancy program, thereby linking high EMT activation to dormancy, already at the level of the primary tumor. Finally, we reveal that the combination of invasive and proliferative signatures is a promising strategy for patient stratification.

Methods

Breast cancer tissue microarray (TMA) analysis

A total of 94 infiltrating grade 3 ductal female breast carcinoma (IDC) samples were obtained from the archives of the MD Anderson Cancer Center, Madrid, Spain (MD Anderson Foundation Biobank, record number B.0000745, ISCIII National Biobank). Patients underwent surgery between 2010 and 2011, with a mean age at surgery of 57.07 years (range: 44–81 years). According to the TNM Classification staging, 32 tumors were stage I, 31 stage II, and 31 stage III–IV. Histological and immunohistochemical studies were performed on formalin-fixed, paraffin-embedded tissue samples[57]. Two 1-mm cylinders were taken from each tumor block to generate TMA blocks[58]. Inter-observer concordance between pathologists was high, exceeding 90%, in line with standard practice for immunohistochemical evaluation in breast pathology. All cases were independently assessed by two experienced pathologists. In instances where substantial discordance was observed, the immunostaining was repeated and evaluated on full tissue sections rather than on TMA cores, allowing for comprehensive reassessment and consensus scoring. The study followed standard ethical procedures outlined in Spanish regulations (Ley de Investigación Orgánica Biomédica, July 14, 2007, and Real Decreto Biobancos 1716/2011) and was approved by the ethics committee of the MD Anderson Cancer Center, Madrid, Spain.

Animal experiments

Mice were fed *ad libitum*. Housing and experimental procedures were conducted in strict compliance with the European Community Council Directive (89/609/EEC) and Spanish legislation. Ethical protocols were approved by the CSIC Ethical Committee and the Animal Welfare Committee of the Institute of Neurosciences. Animals for experiments were selected by genotype, and no randomization or blinding was performed. Animals were housed under SPF conditions at the RMG Animal House (ES-03-119-0001001 SEARMG). Mouse experiments were performed in the MMTV -PYMT model crossed with a *Rosa*-LSL-tdTomato reporter line, purchased from JAX Mice (The Jackson Laboratory), expressing tdTomato upon Cre-mediated recombination. Cre recombinase is expressed under the control of the *KRT14* promoter (Tg(*KRT14-cre*)1Amc/J; 004782). Mice were backcrossed in the FVB background for at least ten generations (99.9% FVB) The generated line was further crossed with *Prrx1* conditional mutant mice (*Prrx1*^{em1An})[19] to give rise to heterozygous (Tg(MMTV-PyMT)634Mul; Tg(KRT14-cre)1Amc; *Prrx1*^{em1An} /+ named as *Prrx1* +/f) and homozygous (Tg(MMTV-PyMT)634Mul; Tg(KRT14-cre)1Amc; *Prrx1*^{em1An} named as *Prrx1* f/f) conditional *Prrx1* mutants. Considering that the study focuses on breast cancer, only female PyMT mice were used. Mammary gland carcinomas were collected from 14 to 16-week-old mice, upon reaching the maximum tumor size or burden permitted by ethical guidelines. Maximal tumor size/burden was defined as equal to or greater than 1500 mm^3 per tumor or equal to or greater than 3000 mm^3 total tumor burden following the ethical protocols approved by the CSIC Ethical Committee and the Animal Welfare Committee of the Institute of Neurosciences (protocols 2015/VSC/PEA/00211 and 2019/VSC/PEA/0218). Maximum size was never exceeded. Health state, tumor size, and burden were monitored weekly.

Immunofluorescence (IF)

Tumor and lung sections. Sample preparation and imaging: 3–5 μm paraffin-embedded sections were dewaxed, and protein epitopes unmasked by immersion in 95 °C preheated citrate buffer (pH 6.0) or Tris-EDTA buffer (pH 9.0) for 20 min. O.C.T.™ sections (5 μm) were washed three times in PBS for 5 min. For IF staining, slides were placed in a humidified chamber and blocked/permeabilized for 1 h with IF blocking buffer (5% normal goat or donkey serum, 1% BSA, and 0.2% Triton X-100 in PBS). Blocking solution was replaced with the primary

antibody diluted in cold IFBB, and the chamber was incubated overnight at 4 °C. Slides were washed 3× for 10 min in PBS, incubated for 1 h with secondary antibodies and DAPI, then washed again 3× for 10 min in PBS. Slides were mounted using an anti-fade mounting medium (DAKO). Primary and secondary antibodies used are shown in Supplementary Table 1. Images were captured using a Leica DMR microscope for brightfield images and an Olympus FV1200 microscope for confocal imaging.

Whole lungs. TdTomato+ metastases were visualized by IF on whole lungs and cleared following the iDISCO+ protocol[59]. Images were acquired using an UltraMicroscope II (LaVision BioTec), analyzed, and reconstructed in 3D using Vision4D (Arivis) Image Analysis Software. Metastatic burden was assessed using Imaris software (version 9.3.1; BitPlane). The "Surface" function was used for segmentation (tdTomato signal), and volumetric data were extracted. To avoid false positives, a detection cutoff of 90,000 μm^3 was identified as the minimum volume for high-confidence object identification.

Mammary tumor and lung sample collection and preparation

Tumor and whole lung samples were fixed in 4% PFA overnight at room temperature (RT) for embedding in paraffin or for 4 h at 4 °C for embedding in O.C.T.™ (Sakura). The next day, samples were washed several times and stored in PBS + Azide overnight at 4 °C. Pre-fixed samples were embedded in paraffin or O.C.T.™ for further sectioning and collection on SuperFrost Plus microscope slides.

Cancer cell FACS sorting and analysis

Mammary gland carcinomas were collected from 14 to 16-week-old female mice. Harvested tissues were manually minced using sterile scalpels and finely chopped with a McIlwain Tissue Chopper (TED PELLA, INC). Minced samples were incubated in 2.5 ml of digestion buffer (PBS supplemented with 1 Wünsch units of TH Liberase/ml and 25 µg/ml DNase I (Roche)) at 37 °C for 75 min in an Incubator Microplate Shaker (VWR). Tumor fragments were pipetted up and down every 5 min to achieve tissue dissociation into single-cell suspensions, and disaggregation was monitored under a microscope. Disaggregated samples were neutralized with 15 mL of breast medium supplemented with 25 µg/ml DNase I and passed through 70 µm and 40 µm filters (BD Falcon). Cells were centrifuged, and pellets were resuspended in 5 mL of red blood cell lysis ACK buffer (Ammonium-Chloride Potassium) for 5 min at RT with gentle rotation. After red blood cell removal, cancer cells were neutralized, resuspended in 0.5 ml FACS buffer with DAPI per 1 M cells, and directly used for FACS to sort Tomato$^+$ cells. Cell sorting was performed using a BD FACSAria III flow cytometer, and 300,000 cancer cells were sorted. The sorted cells were centrifuged at 5000 rpm for 3 min and lysed using the lysis buffer from the *illustra* RNAspin Mini Kit (GE Healthcare) following the manufacturer's instructions. For immunolabeling, after neutralization of the red blood cell lysis buffer, cells were resuspended in 200 µl of FACS buffer per 1 M cells. Fc receptor blocking was performed by adding Palex-BioLegend 156604 (1:50) and incubating for 15 min under agitation at 4 °C. Fluorescence-labeled antibodies were then added, and cells were further incubated for 30 min under agitation at 4 °C. Finally, cells were resuspended in 200 µl of FACS buffer with DAPI per 1 M cells and analyzed using the BD FACSAria III flow cytometer and BD FACSDiva (BD Bioscience) software.

Total RNA extraction, cDNA synthesis, and RT-qPCR

For gene expression analysis, RNA was extracted using the Illustra RNAspin Mini Kit (GE Healthcare). Reverse transcription was performed using the Maxima First-Strand cDNA Synthesis Kit (Thermo Fisher Scientific). RT-qPCR was carried out using Fast SYBR Green Mastermix (Applied Biosystems) on a Step One Plus machine (Applied Biosystems), following the manufacturer's instructions. Relative RNA expression levels (relative fold change) were calculated using the $2^{-\Delta\Delta Ct}$ formula. Quantitative RT-qPCR primers were ordered from IDT, and sequences are listed in Supplementary Table 2.

Patient clustering based on PRRX1 expression

To quantify PRRX1 expression in patients, we computed two metrics for each tumor sample: the percentage and the mean intensity of PRRX1 expression in cancer cells. These values were extracted from single-cell segmentation of IF data using CellProfiler and CellAnalyst softwares[60,61]. The resulting two-dimensional feature space was standardized using StandardScaler from the sklearn.preprocessing module. Unsupervised clustering was then performed using the K-means algorithm (KMeans from sklearn.cluster). The resulting cluster labels were assigned to each patient and used to color-code data points in the scatter plot.

Spatial transcriptomics analysis

Spatial transcriptomics panel design. We designed a MERSCOPE panel of 265 genes (259 genes in the main panel and 6 in the sequential panel) using the Vizgen Gene Panel Design Portal. The genes included are shown in Supplementary Table 3. The gene selection was based on two criteria: Firstly, genes that cover the main cell types and cell states (primarily cancer cell states) were identified in previously published scRNA-seq data of PyMT tumors[19]. Secondly, genes involved in general signaling pathways such as Wnt, Hedgehog, and Hippo pathways, plus cytokine-cytokine receptor interactions, prioritizing those detected in the scRNA-seq data mentioned above. To avoid optical crowding during transcript decoding, we analyzed public bulk RNA-sequencing data from PyMT tumors, and candidate genes with very high expression levels were excluded. For each gene, a panel of 30 encoding probes was designed by Vizgen using a proprietary algorithm. The final gene panel also included 30 blank probes serving as negative controls.

Spatial transcriptomics sample preparation and data acquisition. Formalin-fixed paraffin-embedded (FFPE) tissue from 3 *Prrx1* +/+ PyMT primary tumor was sectioned to a thickness of 5 µm following standard histology protocols. Sections were floated for 5 min in a 45 °C water bath, mounted on fiducial bead-coated FFPE MERSCOPE slides (Vizgen, #20400100), dried at RT in a fume hood for 1 h, baked at 55 °C for 15 min, dried at RT for 2 h, and stored at −20 °C overnight. Samples were prepared following the Vizgen MERSCOPE® User Guide for FFPE Tissue Sample Preparation, beginning at the cell boundary staining step. To remove lipids and proteins that interfere with imaging, 5 ml of Clearing Premix (Vizgen #20300003) was mixed with 100 µl of Proteinase K for each sample. The samples were incubated at 47 °C in a humidified incubator overnight. Samples in Clearing Solution were photobleached for 3 h to remove background fluorescence using the MERSCOPE Photobleacher (Vizgen #1010003). Following photobleaching, the samples were incubated overnight at 37 °C, with the clearing solution replaced with 200 µL of Clearing Premix containing 50 µL of Proteinase K for an additional 4 h at 37 °C. Finally, the samples were incubated for 48 h with the customized gene panel described above. Samples were stored in Clearing Solution at 37 °C in the incubator prior to imaging for up to 1 week. Imaging was performed on the MERSCOPE platform according to the MERSCOPE Instrument User Guide. Each field of view was imaged in seven 1.5 µm-thick z planes at 60× magnification. Images were processed and decoded to identify RNA spots with xyz coordinates and gene assignments using Vizgen's Merlin software.

Cell segmentation. To improve the MERSCOPE onboard cell segmentation, we re-segmented the data using Cellpose2 v2.2.3[62] implemented in the vizgen post-processing tool (VPT) v1.2.2 framework (https://github.com/Vizgen/vizgen-postprocessing). The Cyto2 model provided with Cellpose2 was used for the segmentation. The z_layers

was set to 3, and the three image channels, i.e., Cellbound2, Cellbound3, and DAPI, were mapped to the red, green, and blue channels, respectively. In the preprocessing, gaussian blur (kernel size = 11) was used to smooth the image boundaries. Intensity was normalized using contrast-limited adaptive histogram equalization with a clip limit of 0.01 and a filter size of 100 × 100 pixels. For the segmentation, DAPI was used as a nuclear channel, and the object diameter threshold was set to 70 pixels to specify the approximate cell size. The diameter threshold was derived from the manual inspection of random areas of MERSCOPE onboard segmentation in cellpose2 visualizer. A flow threshold of 0.95 and a cell probability threshold of −5.5 were applied, with a minimum mask size of 500 pixels to filter the noise. Post-segmentation, polygon smoothing, and cell boundaries simplification were performed with a tolerance of 2 pixels and a smoothing radius of 10 pixels, retaining only polygons with a minimum area of 500 pixels. To harmonize the results, entities were fused with a minimum distance of 1 pixel between them. Finally, the partition-transcripts utility from the VPT framework was used to obtain the cell-by-gene matrix using re-segmented data, and derive-entity-metadata was used to derive the cell metadata, which includes the cell coordinates, volume, and transcript counts for each cell. The generate-segmentation-metrics was used to generate the report, which was manually inspected to ensure the quality of re-segmentation of each sample.

Quality control, filtering, and data integration. The cell-by-gene matrix, along with cell metadata, was imported into the Python environment and converted into an AnnData object using Scanpy v1.9.3[63]. Thresholds for gene expression and cell volume were determined by manually inspecting their distributions across each sample. Cells with fewer than 15 gene expression counts, volumes smaller than 50 μm^3, or larger than 5000 μm^3 were removed from downstream analysis, which resulted in 1.794 million total number of cells. AnnData objects from three different samples were merged, and library normalization was performed using the scanpy.pp.normalize_total function. Then the count matrix was log-transformed using the scanpy.pp.log1p function and scaled using the scanpy.pp.scale function (max_value = 10). Principal component analysis (PCA) was performed on the scaled data using scanpy.tl.pca function with default parameters. Finally, sample-level data integration was carried out using Harmony[64], through scanpy.external.pp.harmony_integrate function.

Dimensionality reduction, clustering, and cell type annotation. The neighborhood graph was constructed using the top 20 Harmony components with the scanpy.pp.neighbors function (n_neighbors = 10). The data were represented in low-dimensional space using Uniform Manifold Approximation and Projection (UMAP) using scanpy.tl.umap function (min_dist = 0.5 and spread = 1.0). The initial round of Leiden clustering was performed using the scanpy.tl.leiden function with a resolution of 0.5. After removing low-quality clusters, a second round of clustering was conducted at a resolution of 1.2, following the same pipeline, starting with re-normalizing the count matrix, which resulted in 18 total number of clusters (1.793 million cells). Differential gene expression analysis was carried out using the Wilcoxon test (scanpy.tl.rank_genes_groups), and the expression of bona fide marker genes was used to annotate the cell clusters. The expression was represented as a dot plot to show their specificity using the scanpy.pl.rank_genes_groups_dotplot function.

Identification of invasive edges MERSCOPE data. The cell metadata, along with cell type annotation, was extracted from the AnnData object and imported in MERSCOPE visualizer. The invasive areas were drawn manually by looking for peripheral tumor edges enriched in *Krt14* expression. Cancer cells were categorized as part of the invasive edge or the bulk of the tumor. The metadata for cancer cells annotated as part of the invasive edge or bulk of the tumor was exported and used for the downstream analysis.

Differential gene expression between cancer cells in the invasive edge and the bulk of the tumor. Differential gene expression analysis was performed to compare the cancer cells located at the invasive edge and the bulk of the tumor. The raw count matrix and associated metadata were imported into the R v.4.3.3 (https://www.r-project.org/) statistical framework and converted into a Seurat (v.5.1.0) object using the CreateSeuratObject function. The cancer cells were subsetted from the Seurat object using the subset function and normalized to the library size using the NormalizeData function. The FindMarkers function was used to identify differentially expressed genes between cancer cells at the invasive edge and in the tumor bulk. The resulting statistics were visualized as a volcano plot using ggplot2. Genes upregulated in the invasive edge were identified as the invasive gene signature. The pathway enrichment analysis was performed for the differentially regulated genes (log2FC > 0.25 and P value < 0.05) using EnrichR[65] against MsigDB and the ARCHS4 TFs database. The same gene set was used to create the STINS.

Single-cell RNA-seq analysis

Single-cell preparation. Mammary gland carcinomas were collected from 14- to 16-week-old female mice. Harvested tissues were manually minced using sterile scalpels and finely chopped with a McIlwain Tissue Chopper (TED PELLA, INC). Minced samples were incubated in 2.5 mL of digestion buffer (PBS supplemented with 2.5 Wünsch units of TH Liberase/mL and 25 µg/ml DNase I (Roche)) at 37 °C for 45 min in an Incubator Microplate Shaker (VWR). Tumor fragments were pipetted up and down every 5 min to achieve tissue dissociation into single-cell suspensions, and disaggregation was monitored under a microscope. Disaggregated samples were neutralized with 15 ml of breast medium supplemented with 25 µg/ml DNase I and passed through 70 µm and 40 µm filters (BD Falcon). Cells were centrifuged, and pellets were resuspended in 5 mL of red blood cell lysis ACK buffer (Ammonium-Chloride Potassium) for 5 min at RT with gentle rotation. Following another centrifugation, pellets were resuspended in 5 mL of FACS buffer/5 mM EDTA and passed through a 20 µm filter. To assess single-cell preparation, 20 µl of cell suspension was mixed with an equal volume of Trypan Blue and incubated for 5 min to evaluate cell number, viability, and debris content. Cells were adjusted to a concentration of 1000 cells/µl with cold FACS buffer and used directly for GEM (Gel Bead-In Emulsions) preparation.

Single-cell GEM and cDNA library preparation. Three *Prrx1* +/+, 3 *Prrx1* +/*f*, and 4 *Prrx1 f/f* single-cell preparations were obtained from ten female mice, with each preparation comprising a mix of mammary gland carcinomas from the same mouse. Individual cell encapsulation for single-cell expression profiling and library preparation was performed using the Chromium Single Cell 3′ Library & Gel Bead Kit v3.1 (10x Genomics) and the Chromium Controller (10x Genomics). Samples were multiplexed using the Chromium Dual Index Kit TT Set A. Pooled libraries were loaded onto an Illumina NextSeq 2000 machine with a 100 bp read length.

Quality control, filtering, and data integration. scRNA-seq libraries for different genotypes were sequenced to recover 69436 total number of cells. The raw data quality was inspected using FastQC v0.11.9 (Babraham Institute). The reads were aligned to the mouse genome assembly GRCm38 (mm10; Ensembl reference annotation) and quantified using the 10X Genomics CellRanger v6.1.1 pipeline[66] with default parameters. The CellRanger output was imported into the R v.4.3.3 statistical environment and converted to a Seurat (v.5.1.0) object using the CreateSeuratObject function. The low-quality cells were detected based on the number of genes (400–4000) and the

percentage of mitochondrial genes (15%). The putative doublets were detected using scDblFinder v1.8.0[67]. The low-quality cells and doublets were removed from the downstream analysis. The high-quality cells (n = 50,386) were then subjected to the library normalization using the NormalizeData function with default parameters, followed by the high variable gene (HVGs) detection. In this analysis, we integrated data from 4 similar stage PyMT tumors *Prrx1* +/+ from[19], which were prepared with 10X single cell 3′ v2 chemistry. The top 4000 HVGs were selected for the *Prrx1 f/f* and *Prrx1 +/f* samples, whereas the top 6000 HVGs were selected for the *Prrx1* +/+ samples. The detection of HVGs was performed separately for each sample. Then the intersection was taken for HVGs calculated for each sample per condition, and finally a non-redundant set of genes was taken from the condition specific intersect and used for the sample integration. The gene expression of these HVGs was scaled using the ScaleData function and subjected to PCA. Finally, the samples were integrated based on the top 30 principal components using Harmony[64] implemented in the RunHarmony Seurat function.

Dimensionality reduction, clustering, and cell type annotation. A Shared Nearest Neighbor (SNN) Graph for the integrated Seurat object was built using FindNeighbors over the top 30 harmony components. The UMAP, a non-linear dimensional reduction, was used for the data visualization calculated over the same top 30 harmony components using the RunUMAP function. The single cells were clustered using an SNN modularity optimization-based clustering algorithm implemented in the FindClusters function (resolution: 0.035), which resulted in 9 clusters. First, the gene counts for all the genes in the RNA assay were Log-Normalized using the NormalizeData function, and the differential gene expression was calculated using FindAllMarkers with default parameters. The expression of bona fide markers for each cluster was represented as a dot plot using the DotPlot function, showing their specificity for a cluster.

Clustering of cancer cell subset. The cancer cell cluster was subsetted and re-clustered to better understand the transcriptional heterogeneity within cancer cells. Sample-specific HVGs were using the workflow described above in "Cell quality control, filtering, and data integration." In the initial round of clustering, we used the top 25 Harmony components with a resolution of 1.0. During this step, clusters with high expression of mitochondrial and ribosomal genes were identified and removed as poor-quality clusters. Subsequently, the cleaned cancer cells were re-integrated based on recalculated HVGs using Harmony. A second round of clustering was performed on the top 25 Harmony components, resulting in the identification of 28 distinct clusters at a resolution of 1.5. To annotate the identified cancer cell clusters, a gene signature scoring approach was applied. Briefly, the top 20 differentially expressed genes for each cancer cell cluster were taken from ref. 19, and the gene signature score per cell was calculated using the AddModuleScore_UCell function from the UCell v2.0.0 package[68]. These scores, combined with known marker genes describing cancer cell states, were used to annotate the clusters. In a similar fashion, the score was calculated for dormancy, inflammation, EMT, PINGs, and spatial-derived invasive gene signature.

Condition-specific differential gene expression analysis. To minimize technical variability introduced by different chemistries, for the condition comparison, we utilized only the *Prrx1* f/f samples generated with the 10X single-cell 3′ v3 chemistry, as the *Prrx1* +/f and *Prrx1 f/f* samples were produced using the same v3 chemistry. This sample composition was used while comparing the gene signatures score across different conditions in clusters of interest. To identify condition-specific differentially expressed genes within target clusters, a pseudobulk approach was employed using the muscat v1.16.0 package[69]. Genes with expression count greater than 5 in more than 3 cells were considered for the comparison. The gene counts for each sample within individual cell states were aggregated using the aggregateData function. Genes with a log2 fold change greater than 0.25 and a p value less than 0.05 were considered as upregulated. These upregulated genes were then used for pathway enrichment analysis against the MsigDB database using EnrichR[65].

Single-cell ATAC-seq analysis

Single-nuclei preparation. Four mammary gland carcinomas from *Prrx1* +/+ PyMT mice were collected from 14- to 16-week-old female mice and manually minced using sterile scalpels. Minced samples were transferred to a tissue homogenizer douncer prefilled with Solution D (20 mM Tris-HCl pH 7.5, 0.1% Tween 20, 0.25 M Sucrose, 25 mM KCl, 5 mM $MgCl_2$) and homogenized 8–10 times with rotation. The lysate was passed through 70 µm and 40 µm filters (BD Falcon) and transferred to a 15 mL tube. After centrifugation at 500 × g for 5 min at 4 °C, the supernatant was removed, and the pellet was resuspended in 4 mL of Solution D. Two milliliters of Optiprep (STEMCELL) were added, and the sample was centrifuged at 1500 × g for 10 min at 4 °C. The pellet was resuspended in Wash Buffer (10 mM Tris-HCl, pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 2% BSA, 0.1% Tween 20, 1 mM DTT) and passed through a 40 µm filter. The sample was transferred to a 1.5 mL tube and centrifuged at 500 × g for 5 min at 4 °C. Nuclei were resuspended in 200 µL of 0.1× lysis buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 2% BSA, 0.1% Tween 20, 1 mM DTT, 0.1% Nonidet P40, 0.01% Digitonin) and incubated on ice for 5 min. Following incubation, 1.3 mL of Wash Buffer was added, and the preparation was passed through a 20 µm filter. After centrifugation at 500 × g for 5 min at 4 °C, nuclei were resuspended in 100 µL of 1× Nuclei Dilution Buffer (Chromium Next GEM Single Cell ATAC Kit v2). Ten microliters of the preparation were mixed with an equal volume of Trypan Blue, and the number of nuclei and viability were determined using the automated Countess II cell counter (Invitrogen).

Nucleus transposition and single-nuclei GEM and DNA library preparation. Four single-nuclei preparations were obtained from four female mice, with each preparation comprising a mix of mammary gland carcinomas from the same mouse. Prepared nuclei suspensions were transposed following the guidelines of the Chromium Next GEM Single Cell ATAC Kit v2 (10x Genomics) and encapsulated using the 10x Chromium Controller. Samples were multiplexed using the Chromium Single Index Kit N. Pooled libraries were loaded onto an Illumina NextSeq 2000 machine with a 50 bp read length.

Quality control, filtering, and integration process. Libraries were sequenced to obtain 887.61 million reads in total. The quality of the sequenced reads was assessed using FastQC v0.11.9 (Babraham Institute). Sequenced samples were processed using the CellRanger-atac v2.1.0 pipeline (10x Genomics) and aligned to the GRCm38 (mm10) mouse reference genome (Ensembl annotation v99). MACS2 v2.2.9 was used for peak calling with default parameters. We detected a total of 20,395 cells with 263,346 non-overlapping peaks. Downstream analyses, such as quality control, filtering, normalization, integration, clustering, and visualization, were performed using the R package Seurat v5.0.1[70] and Signac v1.12.0[71]. The appropriate QC matrices were calculated for each sample separately as per the standard Signac guidelines. The sample-specific putative doublets were predicted using AMULET v1.1[72] and filtered out from the subsequent analysis. We retained only high-quality cells (n = 16546) based on the number of fragments falling in peak regions (>3500), percentage of reads in peaks (>30%), and TSS enrichment (>2). The selected threshold was derived by manual inspection of the data distribution. To integrate the different samples, an integration workflow was followed using high-quality cells. Briefly, the four *Prrx1* +/+ samples were merged using the

merge function provided by Seurat, and then Term-Frequency Inverse-Document-Frequency was calculated for the merged object. FindTopFeatures was used to select the top features with min.cutoff = 20. The Singular Value Decomposition (SVD) was performed on the selected top features. Harmony was used to perform the integration in the SVD space. The first component was not considered for any downstream analysis as it is highly correlated with the sequencing depth in sATAC-seq data[71].

Dimensionality reduction and clustering. An SNN Graph was built based on the harmony components using the FindNeighbors function with the top 25 components, and cells were clustered by an SNN modularity optimization-based clustering algorithm implemented in the FindClusters function with a resolution of 0.15. This resulted in six clusters.

Imputation of gene expression and cluster annotation. To impute the gene expression matrix for sATAC-seq assay, we employed the reference mapping approach described by ref. 73. First, the GeneActivity function was used to calculate gene activity score per cell in the sATAC-seq assay for the highly variable genes used in scRNA-seq data. The gene activity score represents the count of fragments located within the 2 kb upstream region from the TSS of a gene. Next, the canonical correlation analysis-based anchors were identified between the scRNA-seq and sATAC-seq assay using the FindTransferAnchors function. The scRNA-seq assay was used as a reference for transferring the gene expression matrix to the sATAC-seq assay using the TransferData function. The imputed gene expression values were visualized on UMAP using the FeaturePlot function to show the specificity of the markers. Additionally, the genomic snapshots of the open chromatin region around selected markers were visualized using the CoveragePlot function.

Clustering of the cancer cell subset. To explore cancer cell heterogeneity at the open chromatin level, we subset cancer cells based on imputed tdTomato expression (also supported by other cancer cell markers). We re-integrated the data as explained in "Quality control, integration, dimensionality reduction and clustering" of sATAC-seq data analysis. Then, the neighbor searching was performed based on the top 20 harmony components, and the FindClusters function was used with the resolution 1.0 to cluster the cancer cells into 16 distinct clusters. Next, we used a similar gene signature-based approach as described in "Clustering of cancer cell subset" of scRNA-seq data analysis to annotate these clusters. The scores were calculated using imputed gene expression values for in sATAC-seq assay.

Motif enrichment analysis of the invasive cluster. The CORE collection of motifs for the vertebrates deposited in the JASPAR database was accessed using the JASPAR2024 v0.99.6 R package[74]. The motif information was added to the Signac object using the AddMotifs function, and the motif activities were calculated per cell using chromVAR[75] implemented in the RunChromVAR function. Finally, the differential motif activity was calculated between the invasive cluster (cluster 15) and the rest of the cancer cell clusters using FindAllMarkers with mean.fxn = rowMeans and max.cells.per.ident = 500.

Prrx1 binding sites in genes of interest. To check whether Prrx1 has direct binding in the promoter regions of the candidate genes for dormancy, cyclins and cyclin inhibitors in the invasive cluster, first open accessible regions were identified in this cluster (cluster 15) using the AccessiblePeaks function. These accessible regions were then annotated to the nearest gene using the annotatePeak function from the ChIPseeker v1.38.0 package[76]. Promotor peaks associated with cyclins, cyclin inhibitors, and dormancy-related genes were analyzed for the presence of Prrx1 motif (JASPAR Database motif accession: MA0716.2), using the FIMO v5.5.7[77] motif search tool from the MEME suite. Coverage plots for these binding sites were generated using the CoveragePlot function.

In silico analysis of human cancer cell lines

The classification of 72 breast cancer cell lines along the epithelial (E) to mesenchymal (M) axis was taken from[19]. Briefly, the enrichment of epithelial (E) and mesenchymal (M) components was computed using epithelial and mesenchymal gene signatures[27] in bulk gene expression data of breast cancer cell lines[39]. The EMT-TFs were excluded from the Mesenchymal signature to avoid biased correlations in the subsequent analyses. Additional scores such as PING[19], Proliferation (https://reactome.org/content/detail/R-HSA-69278), Dormancy[35], and EMT (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp) signatures, along with expression of cyclins, cyclin inhibitors, dormancy markers, and *PRRX1*, were calculated in a similar fashion.

Cell culture

BT-549 (ATCC) and BT-549 sh *PRRX1*[20] human tumor cell lines were cultured in DMEM (Sigma): F12 HAM (Sigma) media (1:1), supplemented with 10% heat-inactivated FBS (Sigma), 10 μg/ml insulin (Roche), 1% Penicillin-Streptomycin (Sigma), and 1% amphotericin (Sigma). MCDK-NBL2 cells (ATCC) were cultured in DMEM (Sigma) supplemented with 10% heat-inactivated FBS (Sigma), 1% Penicillin-Streptomycin (Sigma), and 1% amphotericin (Sigma). Cells were grown at 37 °C and 5% CO_2, with media replaced every 2–3 days. Cells were passaged when they reached 80–90% confluency. For RNA extraction, 2×10^5 BT-549 cells were seeded in six-well plates in complete media and cultured for 48 h, and 10^5 MDCK-NBL2 cells were seeded in six-well plates in complete media and cultured for 72 h.

Stable Prrx1 overexpression cell lines

Mouse Prrx1 was cloned into the pBABE-neo retroviral vector. Retroviral particles were generated in HEK-293T cells by transfection with Lipofectamine 2000 (Thermo Fisher Scientific) together with the packaging plasmids psPAX2 (RRID:Addgene_12260) and pMD2.G (RRID:Addgene_12259). Viral supernatants were collected 48 h post-transfection and used to infect MDCK cells. Forty-eight hours after infection, MDCK cells expressing either pBABE-neo (empty vector) or pBABE-neo–Prrx1 were selected with G418 (400 μg/mL) for 2 weeks. Stable clonal cell lines were subsequently generated by limiting dilution and expanded for downstream analyses.

Transwell cell migration assay

The top chamber insert (Corning Costar Transwell) was covered with 50 μl of mouse collagen IV (Corning, 50 μg ml^{-1}) and left to dry overnight. The resulting matrix was hydrated with 25 μl of water before seeding the cells. 5×10^4 cells from MDCK-NBL2 clones expressing different levels of *Prrx1* were seeded and allowed to migrate. Cell nuclei at the bottom of the insert were imaged 8 h after seeding and automatically counted using ImageJ.

Growth assay

A total of 5×10^4 MDCK-NBL2 cells from clones expressing different levels of Prrx1 were seeded into 24-well plates in complete growth medium. Plates were placed in the Incucyte live-cell imaging system (Sartorius), and phase-contrast images were acquired every hour for 72 h. For each condition, three technical replicates were included in three independent experiments. Cell growth was quantified using the Incucyte's built-in confluency analysis pipeline.

PRRX1 ChIP-Seq data analysis

The raw data for PRRX1 ChIP-Seq in HS578, MDAMB157, and MDAMB436 cell lines were downloaded from the Gene Expression Omnibus database, deposited under GSE202767 accession id. The raw data were filtered using Trimmomatic v0.39 tool with default

parameters. The obtained high-quality reads were then aligned to the human reference genome (hg38) using bowtie v1.0.0 with default parameters. The aligned reads were sorted using samtools v1.11.0, and Picard MarkDuplicates v2.23.0 (https://broadinstitute.github.io/picard/) was used to identify and mark potential duplicated reads. To generate genomic tracks for PRRX1 ChIP-Seq signal for the selected genes deeptools bamCoverage v3.5.2 was used to convert the bam files into normalized coverage files (BigWig). Finally, the BigWig files were visualized in Integrated Genome Visualizer v2.16.0, and genomic snapshots were taken for the selected genes.

Clinical analysis of gene signature in METABRIC data

The gene expression data for METABRIC cohort[5,24,25] along with the clinical parameters was downloaded from the cBioPortal data. The median of log-normalized values was taken for the proliferation (Cell_Cycle_Mitotic_R-HSA-69278) and PINGs (Pro-Invasive genes) signatures. The patients were categorized based on quartiles, into high (75%) and low (25%) groups, combining proliferation (R-HSA-69278) and invasion scores (PINGs). Finally, GraphPad Prism 9.4.1 was used to perform OS analysis, and p value was calculated using the Gehan-Breslow-Wilcoxon test.

Statistics and reproducibility

All experiments were repeated at least three times, and the nature and number of independent experimental replicates are indicated for each experiment in the figure legends. No data were excluded except in scRNA-seq, scATAC-seq, and spatial transcriptomics analyses. In those, bad quality cells were filtered out based on standard practice, as described in the "Methods." No statistical methods were used to predetermine sample sizes. Sample sizes were empirically determined by similarity to those reported in previous publications from our laboratory using similar in vitro and in vivo experimental models. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment. Statistical analyses were performed in R with RStudio or in GraphPad Prism (GraphPad 9.4.1 software). Precise statistical tests are indicated in the figure legends. P values of >0.05 were considered not statistically significant (indicated as "ns"). Boxplots are presented as means ± SEM. Boxplots show median, upper and lower quartiles; whiskers show upper and lower extremes. For OS curves, p values were obtained by using the Gehan-Breslow-Wilcoxon two-tailed test.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The sequencing data generated in this study have been deposited in the Gene Expression Omnibus (GEO) repository. Source data are provided with this paper. Spatial transcriptomics data that support the findings of this study have been deposited and are publicly accessible at the GEO repository: (GSE291210). scRNA-seq data that support the findings of this study have been deposited and are publicly accessible at the GEO repository: (GSE291211). The snATAC-seq data that support the findings of this study have been deposited and are publicly accessible at the GEO repository: (GSE291199). Source data are provided with this paper.

Code availability

Detailed code pipeline has been deposited at a Zenodo repository: https://doi.org/10.5281/zenodo.17726450.

References

1. Klein, C. A. Cancer progression and the invisible phase of metastatic colonization. *Nat. Rev. Cancer* **20**, 681–694 (2020).
2. Luzzi, K. J. et al. Multistep nature of metastatic inefficiency: dormancy of solitary cells after successful extravasation and limited survival of early micrometastases. *Am. J. Pathol.* **153**, 865–873 (1998).
3. Pan, H. et al. 20-year risks of breast-cancer recurrence after stopping endocrine therapy at 5 years. *N. Engl. J. Med.* **377**, 1836–1846 (2017).
4. Pedersen, R. N. et al. The incidence of breast cancer recurrence 10-32 years after primary diagnosis. *J. Natl. Cancer Inst.* **114**, 391–399 (2022).
5. Rueda, O. M. et al. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature* **567**, 399–404 (2019).
6. Lambert, A. W., Zhang, Y. & Weinberg, R. A. Cell-intrinsic and microenvironmental determinants of metastatic colonization. *Nat. Cell Biol.* **26**, 687–697 (2024).
7. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
8. Reiter, J. G. et al. Minimal functional driver gene heterogeneity among untreated metastases. *Science* **361**, 1033–1037 (2018).
9. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
10. Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328–337 (2013).
11. Jakab, M. et al. Lung endothelium exploits susceptible tumor cell states to instruct metastatic latency. *Nat. Cancer* **5**, 716–730 (2024).
12. Yeung, K. T. & Yang, J. Epithelial-mesenchymal transition in tumor metastasis. *Mol. Oncol.* **11**, 28–39 (2017).
13. Youssef, K. K. & Nieto, M. A. Epithelial–mesenchymal transition in tissue repair and degeneration. *Nat. Rev. Mol. Cell Biol.* 1–20. https://doi.org/10.1038/s41580-024-00733-z (2024).
14. Nieto, M. A., Huang, R. Y.-J., Jackson, R. A. & Thiery, J. P. EMT: 2016. *Cell* **166**, 21–45 (2016).
15. Cui, J. et al. MLL3 loss drives metastasis by promoting a hybrid epithelial-mesenchymal transition state. *Nat. Cell Biol.* **25**, 145–158 (2023).
16. Kröger, C. et al. Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. *Proc. Natl. Acad. Sci. USA* **116**, 7353–7362 (2019).
17. Lüönd, F. et al. Distinct contributions of partial and full EMT to breast cancer malignancy. *Dev. Cell* **56**, 3203–3221.e11 (2021).
18. Pastushenko, I. et al. Identification of the tumour transition states occurring during EMT. *Nature* **556**, 463–468 (2018).
19. Youssef, K. K. et al. Two distinct epithelial-to-mesenchymal transition programs control invasion and inflammation in segregated tumor cell populations. *Nat. Cancer* **5**, 1660–1680 (2024).
20. Ocaña, O. H. et al. Metastatic colonization requires the repression of the epithelial-mesenchymal transition inducer Prrx1. *Cancer Cell* **22**, 709–724 (2012).
21. Ferreres, J. R. et al. PRRX1 silencing is required for metastatic outgrowth in melanoma and is an independent prognostic of reduced survival in patients. *Mol. Oncol.* **18**, 2471–2494 (2024).
22. Calabrese, E. J. & Mattson, M. P. How does hormesis impact biology, toxicology, and medicine? *npj Aging Mech. Dis.* **3**, 1–8 (2017).
23. Mattson, M. P. Hormesis defined. *Ageing Res. Rev.* **7**, 1–7 (2008).
24. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
25. Pereira, B. et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 11479 (2016).
26. Pommier, R. M. et al. Comprehensive characterization of claudin-low breast tumors reflects the impact of the cell-of-origin on cancer evolution. *Nat. Commun.* **11**, 3431 (2020).

27. Taube, J. H. et al. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proc. Natl. Acad. Sci. USA* **107**, 15449–15454 (2010).
28. Guy, C. T., Cardiff, R. D. & Muller, W. J. Induction of mammary tumors by expression of polyomavirus middle T oncogene: a transgenic mouse model for metastatic disease. *Mol. Cell. Biol.* **12**, 954–961 (1992).
29. Attalla, S., Taifour, T., Bui, T. & Muller, W. Insights from transgenic mouse models of PyMT-induced breast cancer: recapitulating human breast cancer progression in vivo. *Oncogene* **40**, 475–491 (2021).
30. Wuidart, A. et al. Early lineage segregation of multipotent embryonic mammary gland progenitors. *Nat. Cell Biol.* **20**, 666–676 (2018).
31. Takano, S. et al. Prrx1 isoform switching regulates pancreatic cancer invasion and metastatic colonization. *Genes Dev.* **30**, 233–247 (2016).
32. Cheung, K. J., Gabrielson, E., Werb, Z. & Ewald, A. J. Collective invasion in breast cancer requires a conserved basal epithelial program. *Cell* **155**, 1639–1651 (2013).
33. Ye, X. et al. Distinct EMT programs control normal mammary stem cells and tumour-initiating cells. *Nature* **525**, 256–260 (2015).
34. Sherr, C. J. & Roberts, J. M. Inhibitors of mammalian G1 cyclin-dependent kinases. *Genes Dev.* **9**, 1149–1163 (1995).
35. Ren, Q. et al. Gene expression predicts dormant metastatic breast cancer cell phenotype. *Breast Cancer Res.* **24**, 10 (2022).
36. Dalla, E. et al. Lung-resident alveolar macrophages regulate the timing of breast cancer metastasis. *Cell* **187**, 6631–6648.e20 (2024).
37. Fane, M. E. et al. Stromal changes in the aged lung induce an emergence from melanoma dormancy. *Nature* **606**, 396–405 (2022).
38. Malanchi, I. et al. Interactions between cancer stem cells and their niche govern metastatic colonization. *Nature* **481**, 85–89 (2012).
39. Klijn, C. et al. A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol.* **33**, 306–312 (2015).
40. Jovanović, B. et al. Heterogeneity and transcriptional drivers of triple-negative breast cancer. *Cell Rep.* **42**, 113564 (2023).
41. Comijn, J. et al. The two-handed E box binding zinc finger protein SIP1 downregulates E-cadherin and induces invasion. *Mol. Cell* **7**, 1267–1278 (2001).
42. Cano, A. et al. The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. *Nat. Cell Biol.* **2**, 76–83 (2000).
43. Nicolás, F. J., Lehmann, K., Warne, P. H., Hill, C. S. & Downward, J. Epithelial to mesenchymal transition in Madin-Darby canine kidney cells is accompanied by down-regulation of Smad3 expression, leading to resistance to transforming growth factor-beta-induced growth arrest. *J. Biol. Chem.* **278**, 3251–3256 (2003).
44. Medici, D., Hay, E. D. & Goodenough, D. A. Cooperation between snail and LEF-1 transcription factors is essential for TGF-beta1-induced epithelial-mesenchymal transition. *Mol. Biol. Cell* **17**, 1871–1879 (2006).
45. Vega, S. et al. Snail blocks the cell cycle and confers resistance to cell death. *Genes Dev.* **18**, 1131–1143 (2004).
46. Simeonov, K. P. et al. Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states. *Cancer Cell* **39**, 1150–1162.e9 (2021).
47. Blohmer, M. et al. Quantifying cell divisions along evolutionary lineages in cancer. *Nat. Genet.* 1–12. https://doi.org/10.1038/s41588-025-02078-5 (2025).
48. Lucas, O. et al. Characterizing the evolutionary dynamics of cancer proliferation in single-cell clones with SPRINTER. *Nat. Genet.* **57**, 103–114 (2025).
49. Blassberg, R. et al. Sox2 levels regulate the chromatin occupancy of WNT mediators in epiblast progenitors responsible for vertebrate body formation. *Nat. Cell Biol.* **24**, 633–644 (2022).
50. Beck, B. et al. Different levels of Twist1 regulate skin tumor initiation, stemness, and progression. *Cell Stem Cell* **16**, 67–79 (2015).
51. Calabrese, E. J. Hormetic mechanisms. *Crit. Rev. Toxicol.* **43**, 580–606 (2013).
52. Calabrese, E. J. & Kozumbo, W. J. The hormetic dose-response mechanism: Nrf2 activation. *Pharmacol. Res.* **167**, 105526 (2021).
53. DeBlasi, J. M. et al. Distinct Nrf2 signaling thresholds mediate lung tumor initiation and progression. *Cancer Res.* **83**, 1953–1967 (2023).
54. Modi, S. et al. Trastuzumab deruxtecan in previously treated HER2-low advanced breast cancer. *N. Engl. J. Med.* **387**, 9–20 (2022).
55. Magbanua, M. J. M. et al. Expanded genomic profiling of circulating tumor cells in metastatic breast cancer patients to assess biomarker status and biology over time (CALGB 40502 and CALGB 40503, Alliance). *Clin. Cancer Res.* **24**, 1486–1499 (2018).
56. Agudo, J. et al. Targeting cancer cell dormancy. *Nat. Rev. Cancer* **24**, 97–104 (2024).
57. Cheang, M. C. U. et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J. Natl. Cancer Inst.* **101**, 736–750 (2009).
58. Abd El-Rehim, D. M. et al. High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. *Int. J. Cancer* **116**, 340–350 (2005).
59. Renier, N. et al. iDISCO: a simple, rapid method to immunolabel large tissue samples for volume imaging. *Cell* **159**, 896–910 (2014).
60. Dao, D. et al. CellProfiler Analyst: interactive data exploration, analysis and classification of large biological image sets. *Bioinformatics* **32**, 3210–3212 (2016).
61. Stirling, D. R. et al. CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinform.* **22**, 433 (2021).
62. Pachitariu, M. & Stringer, C. Cellpose 2.0: How to train your own model. *Nat. Methods* **19**, 1634–1641 (2022).
63. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
64. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
65. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* **14**, 128 (2013).
66. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
67. Germain, P.-L., Lun, A., Garcia Meixide, C., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Research* **10**, 979 (2021).
68. Andreatta, M. & Carmona, S. J. UCell: robust and scalable single-cell gene signature scoring. *Comput. Struct. Biotechnol. J.* **19**, 3796–3798 (2021).
69. Crowell, H. L. et al. muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11**, 6077 (2020).
70. Hao, Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).
71. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
72. Thibodeau, A. et al. AMULET: a novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data. *Genome Biol.* **22**, 252 (2021).
73. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).

74. Rauluseviciute, I. et al. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **52**, D174–D182 (2024).
75. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
76. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
77. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

Acknowledgements

The authors thank S. Vega for help and support in managing cell lines and T. M. Gomez, C. López-Blau, and G. Osuna for technical support, and G. Expósito and V. Villar for support at the imaging facility. The authors also thank A. Caler and Pep Mulet for technical help in the FACS/Omics facility. The authors thank Holger Heyn´s team at CNAG (National Center for Genomic Analysis, Barcelona, Spain), especially Sergio Aguilar, Laura Jiménez, and Helena Crowell, for suggestions on scRNA-seq analysis. The authors thank the MD Anderson Foundation Biobank for providing samples (record number B.0000745, Instituto de Salud Carlos III National Biobank). This work was supported by grants MCI PID2021-125682NB-I00 and PID2024-155973NB-I00 to M.A.N., and PID2022-136854OB-I00 to G.M.-B., all funded by MICIU/AEI/10.13039/501100011033 and by FEDER; UE. Funds were also provided by the AECC Scientific Foundation (FC_AECC PRYGN25NIET) and Generalitat Valenciana (Prometeo 2021/45) to M.A.N., and Instituto de Salud Carlos III (CIBERONC, CB16/12/00295) to G.M.-B. M.A.N. also acknowledges financial support from Centro de Excelencia Severo Ochoa, grant CEX2021-001165-S, funded by MCIN/AEI/10.13039/501100011033, and support from the Scientific Network Conexión Cáncer funded by CSIC, Spain. K.K.Y. held an investigator contract from the AECC Scientific Foundation (Ayudas AECC investigador 2022). N.N. is hired under the Generation D initiative, promoted by Red.es, an organization attached to the Ministry for Digital Transformation and the Civil Service, for the attraction and retention of talent through grants and training contracts, financed by the Recovery, Transformation and Resilience Plan through the European Union's Next Generation funds, and R.J.-C. holds a "Severo Ochoa Excellence Programme" PhD contract (PRE2020-091888).

Author contributions

R.J.C., K.K.Y., and M.A.N. conceived the project and interpreted the data. R.J.C. and M.A.N. wrote the manuscript. R.J.C. performed most of the experiments and analyzed the data, and M.A.N. supervised the whole project. N.N. performed the single-cell RNA-seq, single-cell ATAC-seq, spatial transcriptomics, in silico, and human data analyses. K.K.Y. performed immunofluorescence analysis in human tumors, participated in the preparation of single-cell RNA-seq libraries, supervised mice, quantified tumor growth, and analyzed the data. G.M.-B. provided the human breast cancer tissue microarrays and interpreted the immunofluorescence experiments in human tumors. B.L.S. generated the MDCK stable clones. J.G. contributed to the in vitro experiments. M.A.N. also ensured funding.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-70242-4.

Correspondence and requests for materials should be addressed to M. Angela Nieto.

Peer review information *Nature Communications* thanks William Muller, Masakazu Toi, who co-reviewed with Hiroko Masuda, and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.