ZFP281 drives a mesenchymal-like dormancy program in early disseminated breast cancer cells that prevents metastatic outgrowth in the lung

Ana Rita Nobre1,2,6, Erica Dalla1, Jihong Yang3,4, Xin Huang3, Lena Wullkopf1, Emma Risson1, Pedram Razghandi5,6, Melisa Lopez Anton1, Wei Zheng5,6, Jose A. Seoane7,8, Christina Curtis8, Ephraim Kenigsberg9,10,11, Jianlong Wang3 and Julio A. Aguirre-Ghiso1,5,6,12,13,14,15

Increasing evidence shows that cancer cells can disseminate from early evolved primary lesions much earlier than the classical metastasis models predicted. Here, we reveal at a single-cell resolution that mesenchymal-like (M-like) and pluripotency-like programs coordinate dissemination and a long-lived dormancy program of early disseminated cancer cells (DCCs). The transcription factor ZFP281 induces a permissive state for heterogeneous M-like transcriptional programs, which associate with a dormancy signature and phenotype in vivo. Downregulation of ZFP281 leads to a loss of an invasive, M-like dormancy phenotype and a switch to lung metastatic outgrowth. We also show that FGF2 and TWIST1 induce ZFP281 expression to induce the M-like state, which is linked to CDH1 downregulation and upregulation of CDH11. We found that ZFP281 not only controls the early dissemination of cancer cells but also locks early DCCs in a dormant state by preventing the acquisition of an epithelial-like proliferative program and consequent metastases outgrowth.

Most cancer patients die of metastatic relapse, which frequently occurs years to decades after diagnosis and treatment is initiated by DCCs that can remain clinically dormant for long periods. Cancer dormancy is a major clinical problem. However, our knowledge about how cancer cells remain quiescent while retaining metastasis-initiating capacity is still limited. Additionally, it was believed that cancer cells could disseminate and metastasize only during late stages of progression. However, increasing evidence supports that early DCCs seed organs over long periods of time, starting very early in cancer evolution.

Early dissemination or intraorgan dispersion was reported in patients with several cancer types and in several mouse models. However, it is not resolved whether the time it takes for early DCCs to grow into metastases is controlled by an active program that holds early DCCs in a dormant state before they can initiate slow or fast proliferation and continue to evolve genetically.

We previously found that HER2 and PyMT oncogene signaling, along with tissue resident macrophages, activates a partial epithelial-to-mesenchymal transition (EMT) program, leading to early cancer cell dissemination and long-lived metastatic outgrowth in primary sites. Additionally, HER2+ early DCCs in secondary organs maintain a TWIST1+ and long-lived dormant phenotype that preceded metastasis initiation. Here, we used single-cell RNA sequencing (scRNAseq) to reveal the DCC heterogeneity and plasticity of lung DCCs across the spectrum of mammmary cancer progression. We found that the primed pluripotency transcription factor (TF) ZFP281 is a key regulator of early DCC spread and dormancy. Using both organoid and in vivo models, we show that ZFP281 is induced by FGF2 and TWIST1. TWIST1 induces ZFP281, which in turn maintains expression of the former TF to induce M-like and primed pluripotency-like programs. These programs suppress proliferation in primary sites but allow for efficient dissemination (at least to the lungs). After dissemination, ZFP281 maintains early DCCs in a prolonged growth-arrested dormant state via the induction of the class II cadherin 11 (CDH11). Importantly, we show that even aggressive primary tumor (PT) cells, which show low levels of ZFP281, can be reprogrammed into dormancy and prevented from metastasizing by regaining ZFP281 or CDH11 expression. Our findings have yielded a previously unrecognized mechanism of metastatic dormancy that would have been missed if only advanced primary tumor biology and the classical view of the metastatic cascade were considered.

1Division of Hematology and Oncology, Department of Medicine and Department of Otolaryngology, Department of Oncological Sciences, Black Family Stem Cell Institute, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. 2Abel Salazar Biomedical Sciences Institute, University of Porto, Porto, Portugal. 3Department of Medicine, Columbia Center for Human Development, Columbia University Irving Medical Center, New York, NY, USA. 4Zhang Boli Intelligent Health Innovation Lab, Hangzhou, China. 5Department of Cell Biology, Albert Einstein College of Medicine, Bronx, NY, USA. 6Cancer Dormancy and Tumor Microenvironment Institute, Albert Einstein College of Medicine, Bronx, NY, USA. 7Cancer Computational Biology Group, Vall d’Hebron Institute of Oncology, Barcelona, Spain. 8Department of Medicine and Department of Genetics, Stanford University, Stanford, CA, USA. 9Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. 10The Precision Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. 11Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. 12Montefiore Einstein Cancer Center, Albert Einstein College of Medicine, Bronx, NY, USA. 13Gruss-Lipper Biophotonics Center, Albert Einstein College of Medicine, Bronx, NY, USA. 14Ruth L. and David S. Gottesman Institute for Stem Cell Research and Regenerative Medicine, Albert Einstein College of Medicine, Bronx, NY, USA. 15Institute for Aging Research, Albert Einstein College of Medicine, Bronx, NY, USA. 16Present address: Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. 17E-mail: LobatoCA@mskcc.org; julio.aguirre-ghiso@einsteinmed.edu
Results

An active M-like program in early lesions persists in early DCCs.

We used the MMTV-ErbB2/HER2/Neu mouse breast cancer model\(^1\), where we have reproducibly modeled early dissemination and metastatic colonization (also replicated in the MMTV-HER2-T, -HER2-NDL5-CFP and -PyMT models\(^2,3,4\)). The MMTV-HER2 mouse model provides a long temporal window to study early stages of tumorigenesis and metastatic progression before PT detection (Extended Data Fig. 1a\(^5\)). To understand the gene programs present in early versus late MMTV-HER2 lesions, we performed RNA sequencing (bulk RNAseq) of early lesion (EL) and PT spheres, which recapitulate the in vivo behavior of these lesions\(^6,7\). We identified 4,290 differentially expressed genes (DEGs) in EL versus PT spheres (Fig. 1a and Supplementary Table 1). Transforming growth factor β (TGF-β), extracellular matrix, collagen, focal adhesion, P13K and β1 integrin signaling and pathways associated with EMT, among others\(^8,9,10\), were upregulated gene ontology programs enriched in EL cells. In contrast, the top downregulated gene ontology term was tight junction formation (Extended Data Fig. 1b and Supplementary Table 2). Accordingly, gene set enrichment analysis (GSEA)\(^11,12\) revealed EMT as the most enriched hallmark of EL over PT cells (Fig. 1b and Supplementary Table 3). EL cells were also enriched in ‘mammary luminal down’ and ‘mammary stem cell’ signatures\(^13,14\) (Fig. 1b and Supplementary Table 3). Together, these results suggest that MMTV-HER2+ EL cells activate M-like and basal/stem-like programs, which are subsequently silenced in PT cells that gain an epithelial-like (Ep-like) program.

We next characterized the in vivo heterogeneity of EL and PT cells, as well as early and late lung (EL and LL) DCCs using scRNAseq. We sorted MMTV-HER2+ tumor cells from in vivo EL and PT tissues, as well as lungs (EL and LL) carrying DCCs (Extended Data Fig. 1a), and performed scRNAseq (Extended Data Fig. 1d,e). This approach accounts for a fully autochthonous syngeneic and in vivo tissue microenvironment. Gene expression profiles from 3,686 cells were compared with the profiles obtained from EL versus PT sphere bulk RNAseq (Fig. 1c and Extended Data Fig. 1d,e). Single cells sorted from EL tissues showed enrichment in mesenchymal and stemness-associated genes upregulated in the bulk RNAseq of EL over PT spheres and downregulation of luminal genes found downregulated also in EL spheres (Fig. 1c).

This finding supports that gene expression patterns in EL and PT spheres in vitro share a significantly overlapping gene signature from the cells profiled in vivo. Moreover, both EL DCCs and LL DCCs are enriched in genes upregulated in EL over PT (Fig. 1c; \(P=5.31\times10^{-7}\)), suggesting that DCCs may induce or retain gene expression profiles more similar to EL cells, which display a M-like and stemness-like signature. Unsupervised clustering of scRNAseq data, using a previously described batch-aware algorithm\(^16\), showed that EL and PT cells clustered almost independently. Interestingly, lung DCCs also clustered separately from EL and PT cells; however, a single cluster (9) is uniquely composed of EL cells, EL DCCs and LL DCCs, but not PT cells (Fig. 1d), suggesting that ELs contain a subpopulation of cells that already carry a signature similar to the one found in cells that disseminated and persisted in lungs. DCCs from early and late lungs, although heterogeneous and distinct from EL and PT cells, were always contained in the same transcriptional clusters (7–11) (Fig. 1d), suggesting that DCCs with EL signatures may persist in the late stages. This prevented distinguishing early DCCs from those DCCs populating late lungs, most likely because late lungs carry early DCCs (derived from ELs), PT-derived DCCs and growing metastasis, all coexisting in the lungs. Nonetheless, Ep-like and M-like signatures found in the primary early and late lesions (Fig. 1a and Harper et al.\(^19\)) were also found in lung DCCs, and these cells could be broadly grouped into Ep-like (7–8) and M-like (9–11) clusters (Fig. 1c). Ep-like clusters 7 and 8 shared epithelial signatures more homogeneously, whereas the M-Like clusters 9–11 showed nonoverlapping mesenchymal signatures. Of note, DCCs show a high degree of transcriptional heterogeneity, but few cells display full EMT or mesenchymal-to-epithelial transition (MET) signatures, which led the terms M-like and Ep-like rather than strict categorizations.

Analyses by both fluorescence-activated cell sorting (FACS) (Fig. 1f and Extended Data Fig. 1f) and immunofluorescence (Fig. 1g) revealed that freshly isolated PT cells show high levels of HER2 and a predominant epithelial phenotype, characterized by strong EpCAM expression and noninvasive organoids. In contrast, EL cells showed a broader spectrum of HER2 expression and a mixture of epithelial (41.6% EpCAM\(^+\)) and mesenchymal (17.2% Eng/CD105\(^+\)) cell populations (Fig. 1f, representative plots, and Extended Data Fig. 1f, replicates combined), which correlated with a more invasive phenotype (Fig. 1g and Harper et al.\(^19\)). Endoglin (Eng/CD105) was a mesenchymal marker\(^17\) selected from the scRNAseq data due to its selective upregulation in M-like DCCs (clusters 10 and 11; Fig. 1e). We confirmed that the M-like/Eng\(^+\) and invasive phenotype found in EL cells persists and increases in frequency in early lung DCCs (Fig. 1g, Extended Data Fig. 1f, ~56.13%). In contrast, late lungs presented a smaller population of M-like (~30% Eng\(^+\)) and invasive phenotype and increase in Ep-like (~30% EpCAM\(^+\)) and noninvasive phenotype, resembling PT cells (Fig. 1g and Extended Data Fig. 1f). We conclude that a subpopulation of EL cells express an M-like signature found in lung DCCs clustered with an M-like signature. Remarkably, late lungs are still populated by a significant fraction of DCCs with signatures found in early DCCs, similar to EL cells.

Early DCCs gain M-like states and a dormant phenotype.

To gain further insight into the heterogeneity of M- and Ep-like phenotypes of lung DCCs, we next performed additional scRNAseq profiling exclusively on lung HER2\(^+\) DCCs. HER2\(^+\) non-cancer lung cells and HER2\(^+\) DCCs from early- and late-stage mice (15,287\ additional
cells) underwent comprehensive analysis and clustering (Extended Data Fig. 2a–c). Note that the oncogenic driver (for example, HER2) expressed downstream of the MMTV promoter serves as the tag for isolation, and DCCs low or negative for HER2 may be lost and were not considered in the analysis. We identified 25 distinct clusters; 10 were excluded due to their high prevalence in normal lung cells (Fig. 2b). The DCCs were further subgrouped in HER2+ M-like (1 to 4), hybrid (5 to 8) and Ep-like (9 to 15) based on canonical mesenchymal and epithelial signatures (Fig. 2a–c and Extended Data Fig. 2d). The vast majority of DCCs keep an epithelial identity, gaining or losing different degrees of mesenchymal traits (Fig. 2a and Extended Data Fig. 2d). Importantly, el DCCs are enriched in M- and hybrid-like signatures, whereas LL DCCs are more likely to have Ep-like signatures (Fig. 2a,b; chi-squared test, $P < 10^{-10}$).
An extreme Ep-like signature enrichment is cluster 15, which was enriched exclusively in LL DCCs (Fig. 2a). These data suggest that late-stage mice carry more Ep-like DCCs, whereas early-stage mice more frequently have DCCs with M-like and hybrid phenotypes. A caveat of this analysis is that the M- and hybrid-like signatures in LL DCCs may be contributed by eL DCCs that persist in lungs (Fig. 2a,b).

Analysis of gene-to-gene correlation among highly variable genes identified gene modules with strong coexpression patterns (Fig. 2c and Extended Data Fig. 3a). The enrichment of TF targets that correlated with the expression of the modules (Enrichr analysis33,34) revealed multiple programs activated in DCCs that are associated with pluripotency, mixed-lineage differentiation and EMT (Fig. 2d and Supplementary Table 5). Predicted TF enrichment motifs in M-like, hybrid and Ep-like cells were further confirmed by additional differential gene expression analysis between these DCC clusters (Supplementary Table 6). M-like DCCs from cluster 1, enriched in gene module A, express brain- and osteoblast-lineage genes, and this A signature revealed genes that Enrichr analysis identified as being regulated by the TFs Neurod1 (neurogenic differentiation 1), SOX8, SOX9 and SOX10 (embryonic development); and upregulation of Vim, Col1a1 and Col4a2 (EMT genes confirmed by DEG analysis of scRNAseq dataset comparing M-like, hybrid and Ep-like clusters; Supplementary Table 6 and Extended Data Fig. 3). M-like DCCs from cluster 2, enriched in gene module B, share genes mentioned above and genes that the Enrichr analysis identified as regulated by TFs and chromatin remodelers Suz12, Sox17, Sox18 and Pou5f1/Oct4 (pluripotency regulators confirmed by DEG analysis of scRNAseq clusters; Supplementary Table 6). M-like DCCs from clusters 3 and 4 still carried genes controlled by the above TFs but also upregulated EMT genes (Zeb2 and Col3a1) and genes regulated by Snai2, Twist1, Prrx1, Fbn1 (EMT inducers confirmed by DEG analysis of scRNAseq clusters), Supplementary Table 6) and SMADs. These data support the notion that an M-like program initiated in EL cells (Fig. 1 and Extended Data Fig. 1) is also expressed in eL DCCs and persists in LL DCCs. Hybrid DCCs (clusters 5 to 8), shifted toward genes regulated by Gata6, Tp63, Tp73 and Klf4, typical basal and luminal epithelium switch regulators, and epithelial markers Krt7 and Krt8 (Fig. 2c and Supplementary Tables 5 and 6). Thus, DCCs in hybrid clusters might be in transit between M- and Ep-like states. Cluster 8 (hybrid) is composed of distinct cell populations that express gene modules H (B, C and D) and I (Fig. 2e and Extended Data Fig. 2e,f). These DCCs spread between intermediate Ep- and M-like states, starting to lose gene modules B and D while gaining gene module I (Fig. 2e and Extended Data Fig. 2e,f). Gene module I is homogeneously expressed by almost all Ep-like clusters (9 to 14) that expresses epithelial markers confirmed by DEG analysis (Supplementary Table 6 and Extended Data Fig. 3b). Cluster 15 is distinct and, as mentioned above, composed only of DCCs from late lungs (LL). These DCCs express luminal epithelial genes (EpCam and Krt18), Ovol1, Ovol2, Grhl2 TFs (epithelial genes confirmed by DEG analysis – Supplementary Table 6), as well as lactation genes (Csn1s1, Csn2 and Csn3) (Fig. 2b and Supplementary Tables 5 and 6). These findings suggest that cluster 15 corresponds to more differentiated luminal Ep-like DCCs. Importantly, M-like and hybrid DCCs are enriched in dormancy genes, including Nrr2f1, Tgfβ2, Cdkn1c, Wnt5a and Col3a1, among 19 other dormancy-linked genes (hypergeometric test; P < 0.02, Fig. 2f, Extended Data Fig. 3b and Supplementary Table 6). These data indicate that early DCCs activate gene programs of progenitor-, M-like and dormancy phenotypes and that the transition to an Ep-like program is associated with their ability to form proliferative metastasis. Thus, the M- versus Ep-like states reflects a dormant versus proliferative state of DCCs.

ZFP281 is associated with M-like states in ELs and early DCCs. MMTV-HER2 EL cells do not form tumors but disseminate efficiently and persist as DCCs in lungs35. We hypothesized that the M-like program found in the DCCs may be transcriptionally encoded in ELs. To test this hypothesis, we performed a TF network analysis mining the bulk RNAseq data derived from EL versus PT spheres (from Fig. 1). This analysis identified eight interconnected nodes where ZFP281 was the TF node with the highest number of connected DEGs in EL cells (Fig. 3a and Supplementary Table 7). ZFP281 is a key regulator of primed pluripotency in mouse and human embryonic stem cells and functions as a barrier to achieve naïve pluripotency36, and we found only low expression in rare cells in normal mammary gland cells (Extended Data Fig. 4b). Further, ZFP281 promotes EMT in colorectal cancer cells37 and it is upregulated during the naïve-to-primed pluripotent state transition38 where partial EMT/epithelial plasticity was postulated to happen39,40. However, it is unclear how ZFP281 regulates EMT and how it is linked to early breast cancer progression. The second largest node, Nrasa2/Lrh1-1 (Fig. 3a), regulates embryonic stem cell pluripotency39, but its link to EMT in breast cancer is unclear41,42. Among other TFs, RARβ and RARG, previously linked to dormancy35, may also play a role in ELs and early DCCs.

We focused on ZFP281, as in embryogenesis, it regulates stem cell pluripotency, growth arrest and EMT genes. We validated the increase in ZFP281 in EL over PT cells and its computational-predicted target genes (Fig. 3a) levels by quantitative polymerase chain reaction (qPCR) (Extended Data Fig. 4a). We found that M-like genes such as Cdh11 and Eng are induced in EL over PT cells, whereas Ep-like genes such as Cdh11 and EpCAM are downregulated, arguing that ZFP281 represses an Ep-like identity. Predicted ZFP281 target genes (Fig. 3a) are frequently upregulated in M-like lung DCCs, whereas Ep-like lung DCCs do not express these genes (scRNAseq; Fig. 3b). ZFP281 expression itself is also higher in M-like and hybrid cluster cells (1–8) over Ep-like clusters (9–15) (Fig. 3c). At the protein level, we found even greater differences in ZFP281 expression; in normal Friend leukemia virus B susceptible strain (FvB) mouse mammary glands, ZFP281 is
expressed only in 3% of the cells, whereas 30% of MMTV-HER2 EL cells express ZFP281, which is then downregulated in PT cells (8% ZFP281+; Fig. 3d and Extended Data Fig. 4b). The ZFP281+ cells in PTs were mainly found in the tumor–stroma interface. Similarly, in the MMTV-PyMT mouse model, ZFP281 expression is also upregulated in EL over PT cells (Extended Data Fig. 4c). Staining of E-cadherin (Ep-marker) and Twist1 (M-marker) in sequential sections show that in the MMTV-HER2 mouse mammary tissue, HER2+ EL structures enriched in ZFP281 are Ecadlow (less intense membrane staining) and Twist1high (Extended Data Fig. 4b). When monitoring ZFP281 expression in the early HER2+ DCCs, we also found that 42% of single early DCCs are ZFP281+, whereas only 5% of cells within proliferative metastasis are ZFP281+ (Fig. 3e,f). This finding further supports that ZFP281 upregulation in EL cells persist in early lung DCCs. Interestingly, Ki67 and ZFP281 expression were found to be mutually exclusive in early lung DCCs. Interestingly, Ki67 and ZFP281 cells persist in early lung DCCs. Interestingly, Ki67 and ZFP281 expression were found to be mutually exclusive in early lung DCCs (Fig. 3f), suggesting an anti-proliferative function. Further, the majority of lung HER2+ DCCs are in contact with alveolar type II (AT2) cells within the alveoli both in EL (72%) and LL (79%) mice, and all HER2+ DCCs were adjacent (in contact or 1–2 cell diameter) to CD31+ vessels (Extended Data Fig. 4d,e). Among human ductal carcinoma in situ (DCIS) samples, 48% of the cells per lesion were
ZFP281⁺, whereas only 11% of the invasive breast cancer (IBC) cells were positive for ZFP281 (Fig. 3g,h). Detection of Ki67 in these same samples showed that DCIS (ZFP281⁺) was less proliferative than IBC (ZFP281⁻) (Fig. 3i). These data support that ZFP281 is an early breast cancer progression TF, which coordinates EMT-like and growth arrest programs.
Analysis of ZFP281-regulated programs in early DCCs. To reveal the programs that ZFP281 regulates in early mammary cancer cells, we compared RNaseq data from naive versus primed mouse pluripotent stem cells (a transition regulated by ZFP281 (ref. 35)) and MMTV-HER2 EL versus PT spheres (Fig. 1a). Interestingly, DEGs in EL/PT and primed/nairole cells were enriched in hallmarks of EMT (Fig. 1a,b). Conversely, the EMT pathway is downregulated upon RNA interference (RNAi)-mediated ZFP281 downregulation in MMTV-HER2 EL cells (Fig. 4c (right) and Extended Data Fig. 5a), suggesting that in EL cells, ZFP281 may drive EMT. Chromatin immunoprecipitation sequencing (ChIPseq) in EL and PT cells identified 4,018 ZFP281 targets in EL cells (Supplementary Table 8). ZFP281 preferentially binds to 5’ untranslated regions and promoter regions in both EL and PT samples (Extended Data Fig. 5b), and ontology analysis on ZFP281 targets shows EMT, cell cycle and Wnt signaling pathways in the top
enriched pathways (Extended Data Fig. 5c). Strikingly, in EL cells and primed mouse epiblast-derived stem cells (mEpiSCs) (Fig. 4d), ZFP281 seems to regulate overlapping cell cycle arrest, EMT, Wnt and FGFR signaling pathways. Thus, HER2-driven EL cells activate distinct programs found very early in embryo development.

When comparing MMTV-HER2 EL/PT RNAseq and ChIPseq data, we found 759 genes with high ZFP281 binding and high expression in EL cells (Extended Data Fig. 5d and Supplementary Table 10), and gene ontology analysis showed that these genes (UP_UP, red) are enriched in extracellular matrix organization. In contrast, 177 genes have low ZFP281 binding and expression in EL cells (DW_DW, green) and are enriched in cell–cell junction organization, among other pathways (Extended Data Fig. 5e).

Some of these genes overlap with the putative ZFP281 target genes from Fig. 3a (Extended Data Fig. 5f), but we also identified new ZFP281 target genes. Among them are Snai1, Vim, Zeb1 (EMT inducers), Cdk2, Cdkn1a (cell cycle related) and Tgfbr1 and Nr2f1 (dormancy-associated genes) (Fig. 4e, Extended Data Fig. 5g and Supplementary Table 8). These genes were exclusively bound by ZFP281 and upregulated in EL cells or bound by ZFP281 in EL and PT cells but only upregulated in EL cells. We also identified 118 genes with high ZFP281 binding and high expression in PT cells (Extended Data Fig. 5d). This finding suggests that although ZFP281 expression decreases in PT cells, it still binds and regulates a different set of genes of unknown function in PT phenotype. To further filter genes regulated by ZFP281 in EL cells, we compared the list of ZFP281 ChIPseq targets (Supplementary Table 8) with 929 DEGs upon RNAi-mediated ZFP281 downregulation in EL cells (Extended Data Fig. 5a). These data show that more genes are bound by ZFP281 than those detected as induced or repressed directly in EL spheres upon ZFP281 knockdown. Nonetheless, 79 genes are simultaneously bound by ZFP281 in EL cells and differentially expressed upon its downregulation (Fig. 4f).

To address the importance of ZFP281 and its ChIPseq-identified target genes in lung DCCs, we examined their expression in our lung DCC scRNAseq data. Strikingly, M-like and hybrid DCC clusters display the highest levels of ZFP281-regulated signatures (ZFP281 targets from ChIPseq), and these scores drop significantly in Ep-like DCCs (Fig. 4g,h and Extended Data Fig. 5h–j). Some clusters like cluster 8 showed a drop in the ZFP281 signature score, arguing that some hybrid cluster cells move from an M-like to an Ep-like state. Together, the data support that ZFP281-regulated genes are activated in EL cells, carried over and sustained in M-like dormant DCCs in secondary organs.

ZFP281 maintains early DCCs in an M-like dormant state. MMTV-HER2 EL cells are engaged in an M-like invasive program, whereas PT cells have a proliferative phenotype (Harper et al. and Extended Data Fig. 6). To functionally test whether ZFP281 holds DCCs in a dormant state in the lungs, we used an inducible short hairpin for ZFP281 (shZFP281, different targeting sequence from Fig. 5a), confirmed by FACS (Fig. 5d,e) and qPCR (Extended Data Fig. 4a, fourth column), increased sphere formation (Fig. 5b), reduced sphere size and thus proliferation (Fig. 5c) and increased organoid invasive phenotype (Fig. 5f,g). Similarly, overexpression of ZFP281 in MMTV-PyMT PT spheres suppressed sphere size (Extended Data Fig. 6h–j). Thus, PyMT tumors remain responsive to ZFP281 growth-suppressive function when overexpressed.

We next tested the gain- and loss-of-function effects of ZFP281 on tumorigenesis, dissemination and metastasis in vivo. MMTV-HER2 EL spheres transduced with the DOX-inducible shZFP281 system were injected in the mammary fat pad (MFP) and mice were given vehicle drinking water (−DOX), water with doxycycline from day 0 (+DOX), or starting 1 month after sphere injection (−DOX + DOX) for 4 months. As reported previously, few mice developed tumors that were small and static; however, when the injection sites were analyzed after 3 and 5 months, HER2+ EL cells were still found in the MFP of all mice, and DOX treatment caused ZFP281 downregulation (Fig. 6a and Extended Data Fig. 7a). Even in the absence of PTs, after 3 and 5 months (two independent experiments), single DCCs and micrometastases were found in all lungs, supporting a 100% dissemination efficiency by EL cells (Fig. 6b,c and Extended Data Fig. 7b). Three months after downregulation of ZFP281, an increase in the number and area of lung metastasis is already observed compared with shZFP281−DOX mice (Extended Data Fig. 7b). Additionally, both groups of animals in which ZFP281 was downregulated from the beginning (+DOX) or after 1 month (−DOX + DOX) displayed a significant increase in lung metastasis 5 months after injection (Fig. 6c). Importantly, EL shZFP281−DOX + DOX mice showed fewer single lung DCCs than control mice. Although solitary HER2+ DCCs in all groups were Ki67+, proliferative Ki67+ cell frequency in metastasis increased upon ZFP281 downregulation (Extended Data Fig. 7c). Additionally, the frequency of Twist1+ DCCs decreased upon downregulation of ZFP281, whereas Ecad−DCC frequency increased (albeit more variably (not significant)) (Extended Data Fig. 7d,e). Given that the M-like clusters mostly enriched in early DCCs were characterized by a ZFP281-enriched signature that also showed expression of dormancy and cell cycle arrest genes (Fig. 2c,f), these data strongly support that the M-like and dormant phenotypes are induced and maintained by ZFP281. Consistently, loss of ZFP281 signaled early DCC reactivation from dormancy.

Next, we studied the phenotype of PT spheres overexpressing ZFP281 (ZFP281-OE). Control or ZFP281-OE spheres were injected in the MFP, and mice were euthanized 2 or 5 months later (two independent experiments). Tumor sections confirmed an overall increase in ZFP281 expression in the PT ZFP281-OE condition (Fig. 6d and Extended Data Fig. 7i), which resulted in significantly slower growth kinetics, but not tumor take (Fig. 6e), Supporting a growth-suppressive function of ZFP281 (Fig. 5c and Extended Data Fig. 6k). Interestingly, the animals with slower-growing ZFP281-OE tumors (Fig. 6e) showed a fivefold increase in the number of lung single-cell DCCs compared to control tumors (Fig. 6i), but this DCC frequency increase did not result in an increase in micrometastasis at 2 months. Thus, ZFP281 suppresses growth of the PT but enhances dissemination without a subsequent increase in metastatic growth. In a second longer experiment, fewer PT control or ZFP281-OE cells were injected, and tumors were allowed to grow for 70 days and removed by surgery, and then mice were followed and euthanized 5 months after injection. Although no
A significant reduction in the number and size of metastasis was observed in PT ZFP281-OE mice over PT control (Fig. 6g–i), as well as reduction of Ki67+ cells (Extended Data Fig. 7g). Additionally, upon overexpression of ZFP281, the less proliferative DCCs and metastasis displayed higher frequency of TWIST1+ and lower frequency of Ecad+ cells (Extended Data Fig. 7h,i). These results further support the key role of ZFP281 in inducing an M-like phenotype and a growth-arrested dormant phenotype in DCCs.

Upstream inducers and downstream targets of ZFP281. To explore the ZFP281 mechanism of action and regulation, we focused on EMT, Wnt and FGF signaling, common pathways linked to ZFP281 targets in EL and mEpSCs cells (Fig. 4). To this end, we tested several FGF and Wnt ligands (FGF2, FGF10, Wnt3a and Wnt5a) expressed in EL and early DCCs or known to regulate ZFP281 in the embryo. These data revealed that in EL cells, which are already M-like and ZFP281high, only FGF2 could further induce ZFP281 expression (Fig. 7a). Furthermore, TWIST1 and ENG were also induced by FGF2 (Fig. 7a). These results are corroborated by a decrease in the Ep-like population, an increase in the M-like population by FACS (Fig. 7b) and an increase of invasive phenotype of EL cells upon treatment with FGF2 (Fig. 7c). Together, these data suggest that FGF2 is an upstream regulator of ZFP281.

**Fig. 5 | ZFP281 induces an M-like invasive and slow-cycling phenotype in vitro.** a, Representative images of 7-day MMTV-HER2 EL shZFP281±DOX, PT control and PT ZFP281-overexpressed (OE) spheres. Scale bars, 50 μm. b, Quantification of mammosphere (MS) frequency of MMTV-HER2 EL shZFP281±DOX, PT control and PT ZFP281-OE cells. Graph shows n = 3, median and two-tailed Mann–Whitney test. c, Quantification of mammosphere size, as number of cells per sphere after dissociation of MMTV-HER2 EL shZFP281±DOX, PT control and PT ZFP281-OE spheres. Graph shows n = 3, median and two-tailed Mann–Whitney test. d, EpCAM (epithelial marker) and Eng/CD105 (mesenchymal marker) expression in MMTV-HER2 EL shZFP281±DOX, PT control and PT ZFP281-OE spheres. Representative experiment. See panel e for quantification. e, Fold change of Ep-like (EpCAM+Eng−), hybrid (EpCAM+Eng+) and M-like (EpCAM−Eng+) populations in MMTV-HER2 EL shZFP281±DOX over −DOX and PT ZFP281-OE over PT control spheres. Graph shows n = 4, mean, standard error of the mean and two-tailed Mann–Whitney test. f, g, Representative images and quantification of 3D-Matrigel invasive phenotype of MMTV-HER2 EL shZFP281±DOX, PT control and PT ZFP281-OE organoids. Scale bars, 50 μm. Graph shows n = 4, median and two-tailed Mann–Whitney test. See also Extended Data Fig. 6.
Fig. 6 | ZFP281 favors dissemination but serves as a barrier to metastasis initiation by maintaining an M-like dormancy program in early DCCs in vivo.

a, Representative images of HER2 (red) and ZFP281 (green) expression in MFPs of mice injected with MMTV-HER2 EL shZFP281 spheres after 5 months. Mice were given water (1) without doxycycline (−DOX) or (2) with doxycycline for 5 months (+DOX) or (3) 1 month without and 4 months with doxycycline (−DOX +DOX). Arrows, ZFP281+ cells. Scale bars, 25 μm (top row) and 50 μm (bottom row, inserts). Quantification is shown in Extended Data Fig. 7a.

b, Representative images of lung DCCs. HER2, red; Ki67, green; DAPI, blue. Scale bars, 25 μm.

c, Frequency of single cells (SCs) and metastasis 5 months after sphere injections. Two lung slides with all lobules represented were scanned and quantified per mouse. Graph shows n=10 control mice and n=5 mice per remaining conditions, median and two-tailed Mann–Whitney test.

d, Representative images of HER2 (red) and ZFP281 (green) protein expression in PTs 71 days after sphere injections. Scale bars, 25 μm. Quantification in Extended Data Fig. 7f.

e, Frequency of lung SCs and metastasis 2 (f) and 5 (g) months after sphere injections. Two lung slides with all lobules represented were scanned and quantified per mouse. Graph shows n=5 mice per condition, median and two-tailed Mann–Whitney test. See also Extended Data Fig. 7.
TWIST1, a mesenchymal marker, was upregulated in dormant early DCCs, was differentially expressed (mRNA and protein) in early and late DCCs and upon ZFP281 modulation (Fig. 2 and Extended Data Fig. 4a). TWIST1 downregulation in EL spheres using RNAi led to a downregulation of ZFP281, an increase in CDH1 levels (Fig. 7d), as well as a decrease in M-like and invasive phenotypes (Fig. 7e,f). Importantly, these changes were rescued by ZFP281 overexpression (Fig. 7d–f). Our data support a model in which FGFR2 signaling induces ZFP281 to induce TWIST1, which in turn is required to maintain ZFP281 expression to reinforce the M-like phenotype in MMTV-HER2 EL cells.

We next explored ZFP281 downstream mechanisms that might allow early DCCs to maintain an M-like phenotype. CDH11 was consistently upregulated in EL cells downstream of ZFP281, differentially expressed upon ZFP281 modulation and a direct ZFP281 binding target in ChIPseq analysis (Extended Data Figs. 4a,e and 5g). In advanced breast cancer models, CDH11 was found to be upregulated and associated with a mesenchymal phenotype, but no reports link CDH11 to early dissemination and dormancy. Immunofluorescence in the MMTV-HER2 EL and PT lesions, confirmed increased expression of CDH11 in 43% of EL cells versus less than 10% positive cells in PT samples (Extended Data Fig. 8a,b), similar to ZFP281 expression frequency (Fig. 3d). A similar pattern of CDH11 staining and mRNA expression was found in the MMTV-HER2 EL and PT mammosphere cultures (Extended Data Fig. 8c). In testing the functional contribution of CDH11 to EL and PT mammosphere phenotypes, CDH11 downregulation significantly decreased the percentage of invasive EL organoids (Extended Data Fig. 8d,e), whereas overexpression of CDH11 in PT cells led to an increase in organoids with invasive phenotype (Extended Data Fig. 8f,g). This phenocopies the ZFP281 modulation in EL and PT cells, suggesting that ZFP281 may signal through CDH11 (Fig. 4e and Extended Data Fig. 7g).

Similar to ZFP281, CDH11 also seems to have a growth-suppressive function in vivo. CDH11 overexpression in PT cells led to slower tumor growth of orthotopically injected organoids (Extended Data Fig. 8h,i) and significantly reduced outgrowth of lung metastases (Fig. 7g,i). Nevertheless, CDH11 overexpression did not seem to change dissemination to the lungs, as the number of single DCCs and metastasis did not change (Fig. 7h). These data support that CDH11 can to some extent phenocopy the ZFP281 maintenance of an M-like and dormant phenotype.

Discussion
Limited information is available on the fate of the early DCCs between lodging and metastatic outgrowth in target organs. Our publications revealed that oncogene and microenvironmental signals in ELs conspire to activate an EMT program, which persisted in nonproliferative DCCs. Similarly, early pancreatic DCCs also undergo EMT and persist after seeding the liver. Here, we reveal that ELs activate M-like programs linked to primed pluripotency that not only allow EL cells to spread but also enable them with a program of dormancy where stem-cell-like plasticity is operational (Fig. 7j). ZFP281 is expressed in EL cells, enabling dissemination and M-like lung DCCs to explore at least four major transcriptional phenotypes (modules A–D). These M-like programs carry a dormancy signature. Hybrid clusters of DCCs appeared to downregulate ZFP281 activity and regain Ep-like and growth-promoting genes, supporting our hypothesis that ZFP281 prevents DCCs to switch to an Ep-like proliferative phenotype. Interestingly, Ep-like clusters are more homogeneous, arguing that once the DCCs commit to a proliferative phenotype, they are funneled into a more phenotypically uniform state. We interpret that early DCCs enter the lungs in an M-like state and can persist dormant until signals, yet to be determined, cause a final switch. The analysis of CDH11 function suggests that this class II cadherin contributes to maintain the ZFP281-driven M-like program. We hypothesize that an interaction with other CDH11+ DCCs or CDH11+ stromal cells may allow early DCCs to maintain the M-like dormant phenotype. Mechanistic analysis supports that FGFR2 signaling derived from EL cells or other stromal compartments in the primary lesion may enable ZFP281 upregulation for dissemination. Tissue resident macrophages are required for intravasation and early dissemination, but additional work is needed to determine whether they produce FGFR2, or influence the EL or endothelial cells to produce FGFR2 and induce ZFP281 expression. FGFR2 produced by early DCCs or stromal cells in secondary organs may also maintain the dormancy of early DCCs. We also reveal that ZFP281 requires TWIST1 to maintain the M-like phenotype. This finding, together with the finding that other EMT TFs were also upregulated in early DCCs, argues that the M-like program may be quite robust, explaining the ability of these dormant early DCCs to persist long-term. Whether loss of FGFR receptors, TWIST1 or other EMT TFs expressed by early DCCs could also awaken these cells from dormancy remains to be tested.

ZFP281 knockdown switched early DCCs to a more hybrid or Ep-like phenotype, which correlates with increased metastatic reactivation. Consistently, Snail- and Zeb1-driven EMG was previously described to suppress cell cycle progression through repression of cyclin D1 and D2 (refs. 40,41). In contrast, MET was associated with rapid relapse and reduced survival in patients with metastatic castrate-resistant prostate cancer. Further, Lawson et al. found that low-burden (dormant-like) breast cancer DCCs in different organs were mostly basal and pluripotent stem-like, whereas higher-burden DCCs were more luminal-like and proliferative.

We previously reported that the lineage commitment regulators DEC2/BHLHE41 and NRP2F1/COPU-TF1 coordinate stem-like and quiescence programs. However, those studies were in late-evolution cancer models. Our data functionally map these basal/stem-like and developmental/pluripotency programs to such
early stages of cancer evolution and functionally link them with an M-like dormant DCC phenotype. Laughney et al. also reported that metastatic cells (from late-evolution cancer models) recapitulate a primitive transcriptional program spanning stem-like to regenerative pulmonary epithelial progenitor states, such as the key endoderm and lung-specifying TFs SOX2 and SOX9 (ref. 55). Together, these data suggest that pluripotency and dedifferentiation programs may be common in different epithelial cancers and, importantly, already active in early stages of cancer progression. Our findings support that early DCCs display a high degree of cellular plasticity through M-like, primed pluripotency and dormancy programs that likely endow them with the necessary fitness to survive and undergo genetic maturation upon reactivation.

ZFP281 suppresses an epithelial phenotype, inducing a dormant phenotype in early DCCs. Thus, an opportunity opens to identify lesions that may carry or not this dormancy program and determine if it informs on dissemination and relapse. Given that both MMTV-HER2 and MMTV-PyMT models show high expression of ZFP281 in ELs and loss in late lesions, various oncogenic inputs may achieve ZFP281 downregulation with progression. We showed that ZFP281 detection is prevalent in human DCIS samples and significantly decreased in advanced invasive tumors, further supporting
the validity of our findings. ZFP281 seems to be quite specific for ELs and early DCCs. Other TFs, such as NR2F1, which also limits early dissemination, when detected in prostate and breast cancer DCCs inform on patient prognosis. Thus, similar studies could be performed for ZFP281, and its detection may help measure the abundance of early-like DCCs in patients with early or advanced disease and determine whether it serves as a marker of relapse.

More work is needed to validate ZFP281 and the M-like dormancy program in human DCCs, and our approach could not specifically distinguish early DCCs from those exclusively arriving from late lesions. Nevertheless, we provide unprecedented insight into early DCC fate, demonstrating that ZFP281 regulates an active program of dormancy that must be overridden and precedes a slow proliferation phase toward metastasis. Our data may enable exploiting these mechanisms to eliminate DCCs or force them into an indolent and harmless asymptomatic phenotype.

Methods
Animal experiments. All experimental procedures were approved by the Institutional Animal Care and Use Committee of Icahn School of Medicine at Mount Sinai. MMTV-HER2 Neu mice were purchased from the Jackson Laboratory. Four- to 6-week-old female mice were used as early (premalignant) stage mice and 20-week-old or older females with palpable tumors were used as late stage of cancer progression. MMTV-PyMT mice on C57BL/6 background were purchased from the Jackson Laboratory. Four- to 6-week-old female mice were used as early-stage mice and 8-week-old or older females with palpable tumor(s) were used as late stage. No randomization or blinding was used to allocate experimental groups. Tumors were not allowed to grow beyond the size of a 1 cm3 pellet.

Mice were euthanized using isoflurane and cervical dislocation. All five pairs of mammary glands were checked for the presence of any visible small lesions or palpable tumors. Mice were perfused with PBS and organs were collected. For histopathology, organs were fixed in 4% paraformaldehyde (Thermo Fisher Scientific) for 24 h, processed and embedded in paraffin, and sections were cut. For histopathology, organs were fixed in 4% paraformaldehyde (Thermo Fisher Scientific) for 24 h, processed and embedded in paraffin, and sections were cut. Tumors were embedded in paraffin, and sections were cut. H&E slides were scanned using a NanoZoomer S60 Digital slide scanner and NDPview2 software (Hamamatsu), and metastasis area was calculated and normalized for the total area of the lungs.

RNAseq. Total RNA from MMTV-HER2 EL and PT samples (after 7 days in cultures) was extracted using RNeasy protocol (Qiagen) and sequenced using Illumina MiSeq. RNAseq data were analyzed using Basepair software (https://www.basepairtech.com/) with a pipeline that included the following steps. Reads were aligned to the transcriptome derived from UCSC genome assembly ((hg19)) using STAR® with default parameters. Read counts for each transcript was measured using featureCounts®. DEGs were determined using DESeq2 (ref. 4), and a cutoff of 0.05 on adjusted P value (corrected for multiple hypothesis testing) was used for creating lists and heatmaps. GSEA was performed on normalized gene expression counts, using gene permutations for calculating P value (public Gene Expression Omnibus records: GSE165431).

Total RNA was extracted from MMTV-HER2 EL siControl and siZfp281 using Qiagen RNeasy kit. The qPCR was analyzed with an Applied Biosystems 7500 real-time PCR system. The relative gene expression levels were calculated using the 2**(-ΔΔCt) method. For qPCR, forward (F) and reverse (R) primers were used as follows: ZFP281 F: 5′-TGGAAGACCTGCTGTTCTGAC-3′, R: 5′-AAGTCGACCTGCTTGGC-3′; β-actin F: 5′-ATGGCTGTTGACCAAGGAGT-3′, R: 5′-TCTTCTGTTGGCATGGACAT-3′.

Flow cytometry and cell sorting. Mammary glands of early-stage mice, PTs from late-stage mice and lungs from early and late-stage mice were dissected and digested (see Animal experiments section). For the first scRNAseq experiment (Fig. 1), EL, PT and early and late DCCs were sorted (CD45−HER2+), whereas for the 2nd experiment (Fig. 1) non-cancer PTs were sorted on FvB + HER2− and cancer PTs (CD45+ HER2+) were sorted. After sorting, cells were encapsulated using the 10X Chromium v3 kit (first experiment) or v3 (second experiment) and chemistry kit according to manufacturer instructions. Sequencing libraries were prepared according to manufacturer instructions. Quality control of cDNA libraries was performed using CyberGreen qPCR library quantification assay (KAPA). Single-cell RNA sequencing (scRNAseq) was performed on the NextSeq 5500 sequencer (Illumina) paired-end 150 nt read length. RNAseq reads were aligned to the mouse mm10 genome using Bowtie2 (v2.3.4.3). The aligned bam files were sorted by name using the parameter -n. We used HTSeq software (v0.11.2) and the mm10 annotation file from GENCODE (version M19) to count reads for each gene using parameters -r name -f bam and BioMart® to retrieve corresponding gene names. Finally, read counts were normalized with the trimmed mean of M-values method® for differential expression analysis using edgeR® (v3.26.8). RNAseq data were further analyzed using Enrichr® and GSEA®.

Network analysis. Bioinformatic analysis of ZFP281, used in this study, was performed via Enrichr® and GSEA®.

scRNAseq. Mammary glands of early-stage mice, PTs from late-stage mice and lungs from early and late-stage mice were dissected and digested (see Animal experiments section). For the first scRNAseq experiment (Fig. 1), EL, PT and early and late DCCs were sorted (CD45−HER2+), whereas for the 2nd experiment (Fig. 1) non-cancer PTs were sorted on FvB + HER2− and cancer PTs (CD45+ HER2+) were sorted. After sorting, cells were encapsulated using the 10X Chromium v3 kit (first experiment) or v3 (second experiment) and chemistry kit according to manufacturer instructions. Sequencing libraries were prepared according to manufacturer instructions. Quality control of cDNA libraries was performed using CyberGreen qPCR library quantification assay (KAPA). Single-cell RNA sequencing (scRNAseq) was performed on the NextSeq 5500 sequencer (Illumina) paired-end 150 nt read length. RNAseq reads were aligned to the mouse mm10 genome using Bowtie2 (v2.3.4.3). The aligned bam files were sorted by name using the parameter -n. We used HTSeq software (v0.11.2) and the mm10 annotation file from GENCODE (version M19) to count reads for each gene using parameters -r name -f bam and BioMart® to retrieve corresponding gene names. Finally, read counts were normalized with the trimmed mean of M-values method® for differential expression analysis using edgeR® (v3.26.8). RNAseq data were further analyzed using Enrichr® and GSEA®.
digested (see Animal experiments section). In case of cells in culture, single-cell suspensions were obtained by incubating the cells in Accutase (Sigma) for 20 min at 37 °C. Cells were stained using antibodies and conditions in the Reporting summary. All experiments were performed using BD FACSaria II sorter equipped with FACS Diva software (BD Biosciences) or analyzed using Aurora analyzer (Cytek Biosciences) equipped with SpectroFlo software. Dead cells and debris were excluded by FCS, SSC and DAPI (4,6-diamidino-2-phenylindole) (Thermo Fisher Scientific) staining profiles. Data were analyzed with FACS Diva (BD Biosciences) or FCS Express Cytometry 7 (De Novo) software.

**Cell culture.** Mammary glands of early-stage mice, PTs from late-stage mice and lungs from mice at both stages were dissected and digested (see Animal experiments section). For sphere cultures, 5 × 10^5 cells were seeded in six-well ultralow-adhesion plates in 1 ml mammosphere media (DEMEM/F12 (Gibco, 11320–043), 150 ml streptomycin (Gibco, 15050–062), 50 μg/ml hydrocortisone (Lonza CC-403), 40 μg/ml insulin (Gibco 12585–014), 20 ng/ml EGF (Peprotech AF-100–15-A), 100 μM penicillin and 100 μg/ml streptomycin (Corning) supplemented with 0.5% methylcellulose (R&D Systems, HSC001). Sphere-forming capacity was measured by quantification of number of spheres per well after 7 days in culture. Spheres were then dissociated with Accutase (Sigma) and number of cells per sphere was calculated as a measure of sphere size.

EL cells were transduced with lentivirus pTRIPZ (shControl) or shZFP281 (V2THS_42594, Open Biosystems) as previously described67 at day 0 of sphere formation. Cultures were treated every 24 h, starting at day 1, with 2 μg/ml DOX. EL or PT cells were transfected with ZFP281- OE plasmid (p-3ZFP281-OEF (the control vector (Sino Biological, pcMV3-C-GFP-Spark-CV), CDH11-OE (Sino Biological, MG-51164-ACG), siZFP281 (Origene, SR421015), siCDH11 (Ambion, a63755 (A) and a63756 (B)) and siTwist1 (Ambion, 69856) using Lipofectamine 3000 (Invitrogen) or RNAiMax (Thermo Fisher Scientific, 13778) transfection reagents according to the manufacturer's instructions.

For organoid cultures, cells were seeded in eight-well chamber slides coated with 50 μl Matrigel (Corning, growth factor reduced) per slide in 400 s63755 (A) and s63756 (B)) and siTwist1 (Ambion, 69856) using Lipofectamine 3000 (Invitrogen) or RNAiMax (Thermo Fisher Scientific, 13778) transfection reagents according to the manufacturer's instructions.

**Immunofluorescence.** Tissue slides (see Animal experiments section) were dehydrated, followed by antigen retrieval 10 mM citrate buffer, pH 6.0 (Na,H2,H4,O). Blocking was done using 0.5% BSA in PBS with 5% normal goat serum (Thermo Fisher Scientific, PCNS3000) for 1 h. Antibodies and incubation conditions used are summarized in the Reporting summary. For ZFP281 detection, Alexa Fluor 488 Tyramide SuperBoost Kit goat anti-mouse IgG (Invitrogen) was used for amplification of the signal. All slides were mounted with ProLong Gold Antifade reagent with DAPI (Invitrogen, P36931).

3D cultures were fixed with 4% paraformaldehyde for 20 min at room temperature, permeabilized with 0.5% Triton X-100 in PBS for 20 min, and blocking was done using 1% BSA + 2% FCS in PBS for 1 h. Slides were mounted with ProLong Gold Antifade reagent with DAPI (Invitrogen, P36931). Images were obtained using a Leica SPE high-resolution spectral confocal microscope and Leica software.

**qPCR.** Spheres were processed using the Cell-to-C T-1 Step Power SYBR Green kit (Invitrogen, A25600) and primers from Supplementary Table 13. GAPDH or RPLO was used as a housekeeping control for all experiments.

**Patient samples.** Paraffin-embedded sections from DCIS and IBC lesions were collected from the Cancer Biorepository at Icahn School of Medicine at Mount Sinai. Samples were deidentified and obtained with Institutional Review Board approval, which indicated that this work does not meet the definition of human subject research according to the 45 CFR 46 and the Office of Human Subject Research. A total of 28 samples were analyzed, 14 DCIS and 14 IBC.

**Statistical and reproducibility.** No statistical methods were used to predetermine sample sizes, sample sizes were chosen empirically and no exclusion criteria were applied. Block randomization was used for all mice assignments. The investigators were not blinded to allocation during experiments, but calculations were done in coded samples to reduce operator bias. Statistical analyses were done using Prism software, and differences were considered significant if P < 0.05. Exact P values are present in all significant differences; the absence of P values represent nonsignificant differences. Unless otherwise specified, three or more independent experiments were performed in vitro with at least two technical replicates per condition, and two or more independent experiments were performed in vivo with at least five mice per condition. Data distributions were assumed to be normal, but this was not formally tested. Unless otherwise specified, all values were included, median and interquartile range are shown and two-tailed Mann–Whitney U-tests were performed. Representative images were selected after three or more independent experiments were performed and imaged in vitro (Figs. 1 and 5 and Extended Data Fig. 6), and at least five mice per condition were imaged (Figs. 3, 6 and 7 and Extended Data Figs. 7 and 8). Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability.** Source data are provided with this paper. All sequencing data are available in a public data repository (GSE16543), RNAseq of MMTV-Neu EL and PT spheres; GSE16544, ZFP281 ChIPseq of MMTV-Neu EL and PT spheres; GSE165456, scRNAseq of MMTV-Neu primary site and lung cancer cells in early and late stage; GSE165459, scRNAseq of MMTV-Neu lung cancer cells in early and late stage). All other data supporting the findings of this study are available from the corresponding authors on reasonable request.

Received: 9 April 2022; Accepted: 13 July 2022; Published online: 1 September 2022

**References.**

1. Klein, K. C. Cancer progression and the invisible phase of metastatic colonization. Nat. Rev. Cancer 20, 681–694 (2020).
2. Pantel, K. & Brakenhoff, R. H. Dissecting the metastatic cascade. Nat. Rev. Cancer 4, 448 (2004).
3. Valastyan, S. & Weinberg, R. A. Tumor metastasis: molecular insights and evolving paradigms. Cell 147, 275–292 (2011).
4. Yates, L. R. et al. Genomic evolution of breast cancer metastasis and relapse. Cancer Cell 32, 169–184 (2017).
5. Sanger, N. et al. Disseminated tumor cells in the bone marrow of patients with ductal carcinoma in situ. Int. J. Cancer 129, 2522–2526 (2011).
6. Hosseini, H. et al. Early dissemination seeds metastasis in breast cancer. Nature 540, 552–556 (2018).
7. Ullah, I. et al. Evolutionary history of metastatic breast cancer reveals minimal seeding from axillary lymph nodes. J. Clin. Invest. 128, 1355–1370 (2018).
8. Cassendent, A. K. et al. Multiclonal invasion in breast tumors identified by topographic single cell sequencing. Cell 172, 205–217 (2018).
9. Hu, Z. et al. Multi-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastases. Nat. Genet. 52, 701–708 (2020).
10. Rhim, A. D. et al. Detection of circulating pancreas epithelial cells in patients with pancreatic cystic lesions. Gastroenterology 146, 647–651 (2014).
11. Makohon-Moore, A. P. et al. Precancerous neoplastic cells can move through the pancreatic ductal system. Nature 561, 201–205 (2018).
12. Muzumdar, M. D. et al. Clonal dynamics following p53 loss of heterozygosity in Kras-driven cancers. Nat. Commun. 7, 12685 (2016).
13. Shain, A. H. et al. The genetic evolution of metastatic uveal melanoma. Nat. Genet. 51, 1123–1130 (2019).
14. Hu, Z. et al. Quantitative evidence for early metastatic seeding in colorectal cancer. Nat. Genet. 51, 1113–1122 (2019).
15. Turalac, S. et al. Tracking cancer evolution reveals constrained routes to metastases: TRACERx Renal. Cell 173, 581–594 (2018).
16. Li, C. et al. Mutational landscape of primary, metastatic, and recurrent ovarian cancer reveals c-MYC gains as potential target for BET inhibitors. Proc. Natl Acad. Sci. USA 116, 619–624 (2019).
17. Dorssers, L. C. J. et al. Molecular heterogeneity and early metastatic clone selection in testicular germ cell cancer development. Br. J. Cancer 120, 444–452 (2019).
18. Wang, D. et al. Multiregion sequencings reveal the genetic heterogeneity and evolutionary history of osteosarcoma and matched pulmonary metastasis. Cancer Res. 79, 7–20 (2019).
19. Harper, K. L. et al. Mechanism of early dissemination and metastasis in Her2 mammary carcinomas. Cell 588–592 (2016).
20. Rhim, A. D. et al. EMT and dissemination precede pancreatic tumor formation. Cell https://doi.org/10.1016/j.cell.2011.11.025 (2012).
21. Werner-Klein, M. et al. Genetic alterations driving metastatic colony formation are acquired outside of the primary tumour in melanoma. Nat. Commun. 9, 595 (2018).
Extended Data Fig. 1 | Model description, Enrichr analysis, FACS controls and MMTV-HER2 scRNAseq data distribution. (a) Experimental design of MMTV-HER2 bulk and single-cell RNA sequencing experiments. (b) Enrichr analysis of differentially expressed genes (DEG) in MMTV-HER2 early lesion (EL) and primary tumor (PT) 7-day spheres bulk RNAseq. Full table in Supplementary Table 2. Orange, terms mentioned in the text. (c) Biological negative controls used for FACS gating strategy. FvB mammary gland (MG) was used to set the MMTV-HER2 EL and PT gate and FvB lungs for MMTV-HER2 EL and LL DCCs (see Fig. 1f). (d) Number of cells per cluster analyzed in the single-cell RNAseq of MMTV-HER2 EL (teal), PT (red), eL (early lungs, blue) and LL (late lungs, orange) DCCs (see Fig. 1d, e). (e) Number of UMIs per cluster (left) and per sample (right) analyzed in the single-cell RNAseq (see Fig. 1d, e). (f) Percentage of epithelial (EpCAM+Eng−), hybrid (EpCAM+Eng+) and mesenchymal (EpCAM Eng+) populations in CD45−HER2+ MMTV-HER2 EL, PT and eL (early lungs) and LL (late lungs) DCCs after tissue dissociation (representative FACS plots in Fig. 1f). Graph shows n = 5 mice/condition, median, SEM and 2-tailed multiple t-tests.
Extended Data Fig. 2 | MMTV-HER2 scRNASeq data distribution of phenotypes across clusters. (a) Number of cells per cluster analyzed in the single-cell RNAseq of HER2−, HER2+ eL (early lungs) and LL (late lungs) DCCs (see Fig. 2b, c). (b) Number of UMIs per cluster (left) and per sample (right) analyzed in the single-cell RNAseq (see Fig. 2b, c). (c) Scatterplots of single-cell RNAseq datasets (see Fig. 2b, c) using UMAP projections, color coded by per cluster (left) and per sample (right). (d) Distribution of Epithelial (Ep) and Mesenchymal (M) scores (gene lists in Supplementary Table 4, showed in Fig. 2a) in MMTV-HER2 lung DCC clusters. Cell clusters were subgrouped as M-like (1−4, higher M-like score), Hybrid (5−8) and Ep-like (9−15). (e) Distribution of gene modules B and D (M-like) in all DCC clusters. Dots represent single cells color-coded by cluster (left), sample origin (eL or LL, middle) and sub-group (Ep-like, hybrid, M-like, right). Gene module lists in Supplementary Table 4. (f) Distribution of gene modules I (Ep-like) and B (M-like) in all DCC clusters. Dots represent single cells color-coded by cluster (left), sample origin (eL or LL, middle) and sub-group (Ep-like, hybrid, M-like, right). Gene module lists in Supplementary Table 4.
Extended Data Fig. 3 | High resolution plots for scRNASeq clusters and projections on to M-like signatures. (a) Heatmap of UMI counts of selected genes (gene lists in Supplementary Table 4) in MMTV-HER2 eL (early lungs) and LL (late lungs) DCCs single-cell RNAseq after unsupervised clustering on the DEGs and down-sampling to 500 UMI per cell. ‘Per cell’ representation of Fig. 2b heatmap, which shows UMI averages. (b) Single cells color-coded by gene expression and distributed by gene modules B and D (M-like). Examples of EMT- and dormancy-associated genes were selected.
Extended Data Fig. 4 | Functional readouts of ZFP281 gain and loss of function, ZFP281 basal expression in FvB, EL and PT tissues and DCC location in lungs. (a) mRNA expression of ZFP281, its predicted targets (Fig. 3a) and EMT genes in MMTV-HER2 EL versus PT cells, EL shCt, EL shZFP281 and PT ZFP281-OE. Red, upregulated genes; Blue, downregulated genes; two-tailed Mann–Whitney test, \( *p < 0.05 \). (b) Representative images of ZFP281 (1st column, green), E-cadherin (2nd column, green) and Twist1 (3rd column, green) protein expression in consecutive sections of FvB mammary gland (biological negative control) and MMTV-HER2 EL and PT tissues. HER2 expression in red. Arrows point to ZFP281+Ecadlow+Twist1+ cells in EL. Dashed arrow points to ZFP281+ adipocytes (internal control). As previously described, stromal adipocytes express high levels of ZFP281. Scales, 20 μm. (c) ZFP281 expression in MMTV-PyMT EL and PT tissues. Graph shows \( n = 9 \) slides from 5 mice per group, median and two-tailed Mann–Whitney test. (d) Representative images of the location of lung DCCs in relation to alveolar type II (AT2) cells and CD31+ vessels. Scales, 25 μm. (e) Quantification of lung DCCs in contact with alveolar type II (AT2) cells in MMTV-HER2 eL and LL. Graph shows \( n = 3 \) mice per group, median and two-tailed Mann–Whitney test.
Extended Data Fig. 5 | RNAseq and ChIP-seq analysis for ZFP281 targets and projection of ZFP281 binding score on to M-, H- and Ep-like clusters.

(a) Volcano plot from RNAseq data on MMTV-HER2 siZFP281 cells. 436 downregulated genes (green), p-value<0.05 & log2FC < −0.5; 493 upregulated genes (red), p-value<0.05 & log2FC > 0.5. Gene lists in Supplementary Table 10. (b) Distribution of ZFP281 binding peaks localization in both MMTV-HER2 EL (top) and PT (bottom) cells. Graph shows n = 3 and mean. (c) Global analysis on ZFP281 targets in MMTV-HER2 EL versus PT cells. Dotted line, p-value=0.05. (d) Volcano plot of combined RNAseq (x) and ChIPseq (y) from MMTV-HER2 EL versus PT cells. 143 genes show lower ZFP281 binding and higher expression in EL versus PT cells (DW_UP); 759 genes show higher ZFP281 binding and higher expression (UP_UP); 177 genes show lower ZFP281 binding and lower expression (DW_DW); 121 genes show higher ZFP281 binding and lower expression (UP_DW). Gene lists in Supplementary Table 11. (e) Global analysis on genes with high ZFP281 binding and high expression (red, UP_UP) and low ZFP281 binding and low expression (green, DW_DW) in MMTV-HER2 EL vs PT cells (identified in C). (f) Venn diagram of MMTV-HER2 EL versus PT RNAseq (Fig. 1a), ZFP281 node (Fig. 3a) and ChIPseq (Fig. 4b) data. Targets of ZFP281 in EL cells and EpiSCs were identified from ChIP-seq data and further used to compare with EMT, Wnt, FGFR, and cell cycle arrest genes. (g) Representative tracks of MMTV-HER2 EL/PT ChIPseq (Fig. 4b, c). EMT genes: Snai1, Vim, Zeb1, Cdh11, Twi1; Cell cycle associated genes: Cdkn2d, Cdkn1a; Dormancy-associated genes: Tgfbr1, Nr2f1. (h) Frequency of ZFP281 target (ChIP) score, summarizing the averaged expression of ZFP281 targets, in all cells analyzed by scRNAseq (Fig. 2). (i, j) Distribution of ZFP281 target scores and modules D and B (M-like) (i) or modules I (Ep-like) and B (M-like) (j) in all DCC clusters. Dots represent single cells color-coded by ZFP281 target scores (low, red to high, green).
Extended Data Fig. 6 | See next page for caption.
Extended Data Fig. 6 | Modulation of M-, Hybrid and Ep-like phenotypes upon ZFP281 gain and loss of function in 3D cultures and mammospheres.

(a) Representative images of the mammosphere phenotype of MMTV-HER2 EL, PT and EL shControl±DOX cells. Scale 50 μm. (b) Quantification of mammosphere (MS) frequency of MMTV-HER2 EL, PT and EL shControl±DOX cells. Graph shows n = 3, median and two-tailed Mann--Whitney test. (c) Quantification of mammosphere (MS) size, as number of cells per sphere after dissociation of MMTV-HER2 EL, PT and EL shControl±DOX spheres. Graph shows n = 3, median and two-tailed Mann--Whitney test. (d) EpCAM (epithelial marker) and Eng/CD105 (mesenchymal marker) expression in MMTV-HER2 EL, PT and EL shControl±DOX cells. Representative experiment. (e) Fold change of Ep-like (EpCAM + Eng−), hybrid (EpCAM + Eng+) and M-like (EpCAM−Eng+) populations in EL over PT and EL shControl±DOX spheres. Graph shows n = 4, mean, SEM and two-tailed Mann--Whitney test. (f, g) Representative images and quantification of 3D-Matrigel invasive phenotype of MMTV-HER2 EL, PT and EL shControl±DOX organoids. Scale 50 μm. Graph shows n = 4, median and two-tailed Mann--Whitney test. (h) mRNA expression of ZFP281 in MMTV-Neu EL siControl and siZFP281. Graph shows n = 2, and median. (i) Quantification of 3D-Matrigel invasive phenotype of MMTV-HER2 siControl and siZFP281. Graph shows n = 4, median and two-tailed Mann--Whitney test. (j, k) Quantification of mammosphere (MS) frequency (J) and size (as number of cells per sphere after dissociation, K) of MMTV-PyMT EL and PT spheres. Graph shows n = 5 experiments for MMTV-PyMT EL conditions and n = 3 for MMTV-PyMT PT conditions, median and two-tailed Mann--Whitney test. (l) mRNA expression of ZFP281 in MMTV-PyMT EL and PT spheres. Graph shows n = 3, median and two-tailed Mann--Whitney test.
Extended Data Fig. 7 | Dormant versus metastatic fate and M- versus Ep-like phenotypes of DCCs in target organs. (a) ZFP281 expression of MMTV-HER2 EL-shZFP281 -DOX, +DOX and -DOX+DOX cells in the mammary fat pad, 5 month after injection. Graph shows n = 4 mice per condition, median and two-tailed Mann–Whitney test. (b) Frequency of lung metastasis and area, 3 month after MMTV-HER2 EL-shZFP281 sphere injections. Graph shows n = 5 per condition, median and two-tailed Mann–Whitney test. (c) Quantification of Ki67+ cells in lung metastasis 5 months after MMTV-HER2 EL shZFP281 sphere injections. Graph shows n = 5 mice per condition, median and two-tailed Mann–Whitney test. (d, e) Quantification and representative images of Twist1+ and Ecad+ cells in lung metastasis 3 months after MMTV-HER2 EL shZFP281 sphere injections. Graph shows n = 3 mice per condition for Twist quantifications and n = 5 mice per condition for Ecad quantifications, median and two-tailed Mann–Whitney test. Scales 25 μm. (f) ZFP281 expression of MMTV-HER2 PT Control or PT ZFP281-OE primary tumors. Graph shows n = 4 mice per condition, median and two-tailed Mann–Whitney test. (g) Quantification of Ki67+ cells in lung metastasis 5 months after MMTV-HER2 PT Control or PT ZFP281-OE sphere injections. Graph shows n = 5 mice per condition, median and two-tailed Mann–Whitney test. (h, i) Quantification and representative images of Twist1+ and Ecad+ cells in lung metastasis 5 months after MMTV-HER2 PT Control or PT ZFP281-OE sphere injections. Graph shows n = 3 mice per condition for Twist quantifications and n = 5 mice per condition for Ecad quantifications, median and two-tailed Mann–Whitney test. Scales 25 μm.
Extended Data Fig. 8 | Characterization and functional analysis of CDH11 expression in EL and PT lesions. (a, b) CDH11 (green) protein expression in MMTV-HER2 (HER2, red) EL and PT cells. Scales, 25 μm. Graph shows n = 3 mice per group, median and two-tailed Mann–Whitney test. (c) CDH11 mRNA expression in MMTV-HER2 EL and PT spheres. Graph shows n = 3, median and two-tailed Mann–Whitney test. (d, e) mRNA expression of CDH11 (D) and 3D-Matrigel invasive phenotype (E) of MMTV-HER2 EL organoids transfected with siControl or siCDH11. Graphs show n = 2 (D) and n = 4 (E), median and two-tailed Mann–Whitney test. (f, g) mRNA expression of CDH11 (F) and 3D-Matrigel invasive phenotype (G) of MMTV-HER2 PT organoids 7 days after CDH11-OE. Graphs show n = 3 (F) and n = 4 (G), median and two-tailed Mann–Whitney test. (h) Representative images of CDH11 (red or gray) protein expression in primary tumors and lung metastasis of mice injected with PT Control and PT CDH11-OE spheres. Scales 20 μm. (i) Tumor volume over time of PT Control and PT CDH11-OE mice, until the primary tumor reached size for surgery. Graph shows n = 6 PT Control and 8 PT CDH11-OE mice, median, interquartile range and two-tailed multiple t-tests.
Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
  - Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  - Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: No software was used for data collection.

Data analysis: Basepair, Enrichr and GSEA were used for data analysis. Detailed information about the analysis can be found in the Methods section.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The datasets generated during and/or analyzed during the current study are available within the paper, extended data and Source data files. All sequencing data is available in a public data repository. All other data supporting the findings of this study are available from the corresponding authors on reasonable request.

GSE165431, RNA sequencing of MMTV-Neu early lesion (EL) and primary tumor (PT) spheres
GSE165444, ZFP281 ChIP sequencing of MMTV Neu early lesion (EL) and primary tumor (PT) spheres
Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender: Samples were collected from the Cancer Biorepository at Icahn School of Medicine at Mount Sinai, New York, New York. 28 samples were analyzed, 14 carcinoma in situ (DCIS), 14 invasive breast cancer (IBC).

Population characteristics: Women with DCIS or IBC diagnosed, treated and with samples available in the Cancer Biorepository at Icahn School of Medicine at Mount Sinai, New York.

Recruitment: No recruitment of participants was needed. Samples were collected from a biorepository.

Ethics oversight: Samples were de-identified and obtained with Institutional Review Board approval (Program for the Protection of Human Subjects / Institutional Review Board of Icahn School of Medicine at Mount Sinai), which indicated that this work does not meet the definition of human subject research according to the 45 CFR 46 and the Office of Human Subject Research.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☑ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-list.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size: No statistical methods were used to pre-determine sample sizes, sample sizes were chosen empirically.

Data exclusions: No exclusion criteria were applied.

Replication: Unless otherwise specified, 3 or more independent experiments were performed, in vitro with at least 2 technical replicates per condition; and 2 or more independent experiments were performed in vivo with at least 5 mice per condition.

Randomization: Block randomization was used for all mice assignments.

Blinding: The investigators were not blinded to allocation during experiments but all quantifications were done in coded samples to reduce operator bias.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a Involved in the study

☑ Antibodies

☑ Eukaryotic cell lines

☑ Palaeontology and archaeology

☑ Animals and other organisms

☑ Clinical data

☑ Dual use research of concern

Methods

n/a Involved in the study

☑ ChiP-seq

☑ Flow cytometry

☑ MRI-based neuroimaging
Antibodies

Antibodies used
- PE anti-mouse HER2, Novus Biologicals, NBP2-34641PE, FACS: 1:100, 15min at 4°C;
- APC anti-mouse CD45, Biolegend, 103112, FACS: 1:100, 15min at 4°C;
- PerCP/Cy5.5 anti-mouse CD326 (EpCam) Biolegend, 118220, FACS: 1:100, 15min at 4°C;
- PE/Cyanine7 anti-mouse CD105 (Eng) Biolegend, 120410, FACS: 1:100, 15min at 4°C;
- HER2, Millipore, OP15L, IF: 1:100, ON at 4°C;
- HER2, Abcam, ab214275, IF: 1:100, ON at 4°C;
- ZFP281 Santa Cruz, sc-166933, IF: 1:50, ON at 4°C;
- Ki67, Invitrogen, 14-5698-82, IF: 1:100, ON at 4°C;
- E-cadherin, BD Biosciences, 610182, IF: 1:100, ON at 4°C;
- Twist1, Millipore, ABQ29, IF: 1:50, ON at 4°C;
- Coh11, Invitrogen, 71-7600, IF: 1:100, ON at 4°C;
- GFP, Aves, GFP-1020, IF: 1:500, ON at 4°C.

Validation
All antibodies used are commercially available and tested by manufactures. All antibody staining were done in parallel with negative controls to guarantee specificity.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s) Only primary cell cultures were used in this study. Cells were isolated from mice for each experiment and kept in culture no more than 2 weeks.

Authentication No cell line authentication was performed.

Mycoplasma contamination No mycoplasma tests were performed.

Commonly misidentified lines (See ICLAC register) No commonly misidentified cell line was used.

Animals and other research organisms

Policy information about studies involving animals: ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals MMTV-HER2/Neu mice were maintained on FvB background and bred and crossed in our facilities. 14 to 18-week-old female mice were used as early (‘pre-malignant’) stage mice and 20-week-old or older females with palpable tumor(s) were used as late stage of cancer progression. MMTV-PyMT mice on C57/BL6 background were purchased from the Jackson Laboratory. 4 to 6-week-old female mice were used as early stage mice and 8-week-old or older females with palpable tumor(s) were used as late stage. BALB/c nu/nu mice were purchased from Charles River. No randomization or blinding was used to allocate experimental groups. Tumors were not allowed to grow beyond the IACUC allowed limit of 1 cm3 per animal.

Wild animals No wild animals were used in this study.

Reporting on sex Only female mice were used in this study.

Field-collected samples No field-collected samples were used in this study.

Ethics oversight All experimental procedures were approved by the Institutional Animal Care and Use Committee (IACUC) of Icahn School of Medicine at Mount Sinai.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

ChIP-seq

Data deposition
- Confirm that both raw and final processed data have been deposited in a public database such as GEO.
- Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE165444

Files in database submission
- Raw (fastq) and processed (bed) files were submitted for all samples. GSM5033687 Ch IP Early lesion (EL) spheres - input GSM5033688 Ch IP Early lesion (EL) spheres - replicate 1 GSM5033689 Ch IP Early lesion (EL) spheres - replicate 2 GSM5033690 Ch IP Early lesion (EL) spheres - replicate 3 GSM5033691 Ch IP Primary tumor (PT) spheres – input
Methodology

Replicates
One EL input, one PT input, three biological EL and three biological PT replicates ChiPed DNA were used to prepare ZFP281 ChiP-seq libraries.

Sequencing depth
Massively parallel sequencing was performed with the Illuma HiSeq4000 according to the manufacturer protocol, and paired-end 150 bp-length reads were produced.

Antibodies
ChIP was performed using EZ ChIP protocol (Millipore, Massachusetts, USA) with Zfp281 antibody (ab101318, Abcam) for EL and PT samples (after 7 days in cultures).

Peak calling parameters
Chip-seq reads were aligned to the mm10 genome using bowtie2 (v2.3.4.3), followed by removing PCR duplicates using Picard with the parameter REMOVE_DUPLICATES=true. Ch IP-seq peaks were determined by the MACS program (v2.1.2) using input Ch IP-seq as the control data, and all other parameters followed the default setting.

Data quality
Binding difference around the transcription start sites [-5kb, +5kb] between the EL and PT samples are analyzed using the DiffBind (v2.1.6).

Software
Mentioned above.

Flow Cytometry

Plots

Confirm that:
☑ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
☑ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
☑ All plots are contour plots with outliers or pseudocolor plots.
☑ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation
Mammary glands of 'early stage' mice, primary tumors from 'late stage' mice and lungs from 'early' and 'late stage' mice were dissected and digested (see 'Animal experiments' section). In case of cells in culture, single cell suspensions were obtained by incubating the cells in accutase (Sigma) for 20 minutes, at 37°C. Cells were stained using antibodies and conditions described in the Antibodies section.

Instrument
All experiments were performed using BD FACSARia II sorter equipped with FACS Diva software (BD Biosciences) or analyzed using Aurora analyzer (Cytex Biosciences) equipped with SpectroFlo software.

Software
Data were analysed with FACS Diva (BD Biosciences) or FCS Express Cytometry 7 (De Novo) softwares.

Cell population abundance
HER2+ cells vary from 0.5-2% in lungs to 80-90% in primary tumors. Biological negative (corresponding FvB tissues) and positive (primary tumor) were used as fluorescent control and set the sorting gates. 4-way purity was used in all sorting experiments.

Gating strategy
Dead cells and debris were excluded by FCS, SSC and DAPI (4',6-diamino-2-phenylindole) (Fisher Scientific) staining profiles.

☑ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary information.