



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

Ana Rita Nobre^{1,2,16}✉, Erica Dalla¹, Jihong Yang^{1,3,4}, Xin Huang^{1,3}, Lena Wullkopf¹, Emma Risson¹, Pedram Razghandi^{5,6}, Melisa Lopez Anton¹, Wei Zheng^{5,6}, Jose A. Seoane^{1,7,8}, Christina Curtis⁸, Ephraim Kenigsberg^{9,10,11}, Jianlong Wang^{1,3} and Julio A. Aguirre-Ghiso^{1,5,6,12,13,14,15}✉

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 and is initiated by DCCs that can remain clinically dormant for long periods¹. 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^{2–4}. However, increasing evidence supports that early DCCs seed organs over long periods of time, starting very early in cancer evolution¹.

Early dissemination or intraorgan dispersion was reported in patients with several cancer types^{5–19} and in several mouse models^{7,11,20–23}. 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 proliferation¹ 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^{7,20,24}. Additionally, HER2⁺ early

DCCs in secondary organs maintain a TWIST1⁺ and long-lived dormant phenotype that preceded metastasis initiation²⁰. Here, we used single-cell RNA sequencing (scRNAseq) to reveal the DCC heterogeneity and plasticity of lung DCCs across the spectrum of mammary 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. ¹¹Icahn Institute 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. ✉e-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 model²⁵, where we have reproducibly modeled early dissemination and metastatic colonization (also replicated in the MMTV-HER2-T, -HER2-NDL5-CFP and -PyMT models^{20,24}). 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. 1a²⁰). 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 lesions²⁰. 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, PI3K and β 1 integrin signaling and pathways associated with EMT, among others^{26,27}, 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)^{28,29} 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' signatures³⁰ (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; $P = 5.31 \times 10^{-7}$), 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 algorithm³¹, 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.²⁰) 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. 1e). 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.

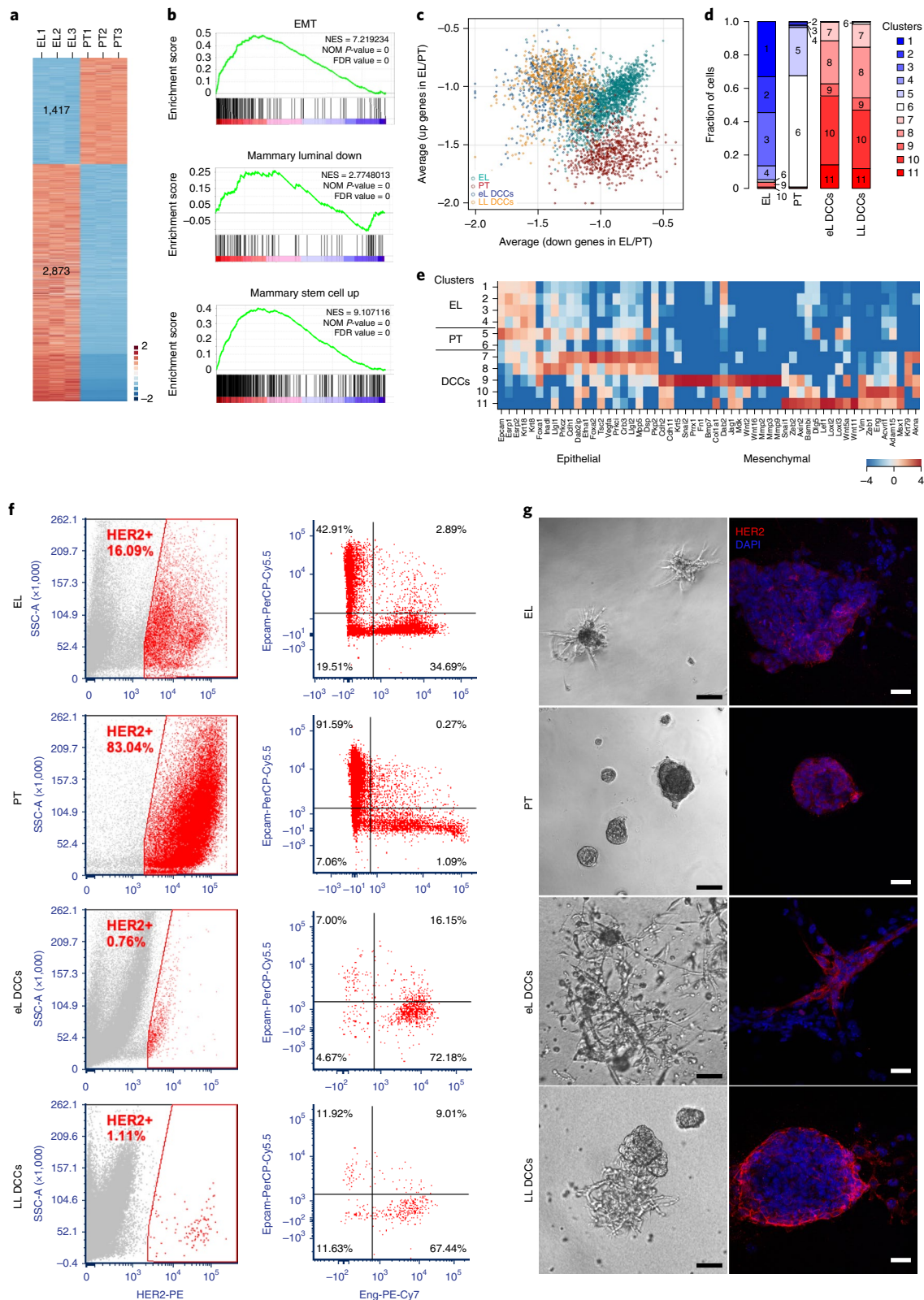
Analyses by both fluorescence-activated cell sorting (FACS) (Fig. 1f and Extended Data Fig. 1f) and immunofluorescence (Fig. 1g) revealed that freshly isolated PT cells show high levels of HER2 and a predominant epithelial phenotype, characterized by strong EpCAM expression and noninvasive organoids. 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.²⁰). Endoglin (Eng/CD105) was a mesenchymal marker³² 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. 1f,g, 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. 1f,g 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

Fig. 1 | Early DCCs maintain a global M-like phenotype. Heatmap of DEGs (Supplementary Table 1) detected by RNAseq from MMTV-HER2 EL and PT spheres cultured for 7 days showing 2,873 upregulated genes (red) and 1,417 downregulated genes (blue) (P value < 0.05 and fold change (FC) >2 or <0.5). Full table is shown in Supplementary Table 1. **a**, GSEA^{28,29} of EMT (top hallmark hit) and mammary gland luminal, myoepithelial/basal and stem cell signatures³⁰ in MMTV-HER2 EL versus PT 7-day sphere bulk RNAseq (GSEA; Supplementary Table 3). ES, enrichment score; NES, normalized enrichment score; NOM P value, nominal P value < 0.05; FDR value, false discovery rate < 0.25. Full table is shown in Supplementary Table 3. **b**, Distribution of the gene expression signatures 'Up' (upregulated) and 'Down' (downregulated) in EL/PT spheres (Fig. 1a and Supplementary Table 1) in MMTV-HER2 EL (teal), PT (red), eL (early lungs, blue) and LL (late lungs, orange) DCC scRNAseq. **c**, MMTV-HER2 EL, PT, eL (early lungs) and LL (late lungs) DCC scRNAseq sample distribution per cluster. Unsupervised clustering on the DEGs was performed using a previously described batch-aware algorithm³¹. **d**, Heatmap of unique molecular identifier (UMI) counts of epithelial (Ep) and mesenchymal (M) genes (Supplementary Table 4) in scRNAseq after unsupervised clustering on the DEGs using a previously described batch-aware algorithm³¹. Cell clusters were subgrouped as EL (1–4), PT (5–6) and DCCs (7–11), according to the predominant cell type in each cluster. **e**, Representative plots of EpCAM (epithelial marker) and Eng (mesenchymal marker) CD45⁻HER2⁺ MMTV-HER2 EL, PT cells and eL (early lungs) and LL (late lungs) DCCs after tissue dissociation. Quantification is shown in Extended Data Fig. 1f. **f**, Imaging of sorted CD45⁻HER2⁺ EL, PT, eL and LL DCCs upon 7-day culture in 3D, on top of Matrigel. SSC, side scatter. **g**, Brightfield images (left; scale bars, 50 μ m) and HER2 (red) expression (right; scale bars, 25 μ m). See also Extended Data Fig. 1.

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 analysis^{33,34}) 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*, *Col4a1* 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 *NR2F1*, *TGFβ2*, *CDKN1C*, *Wnt5a* and *Col3a*, 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.

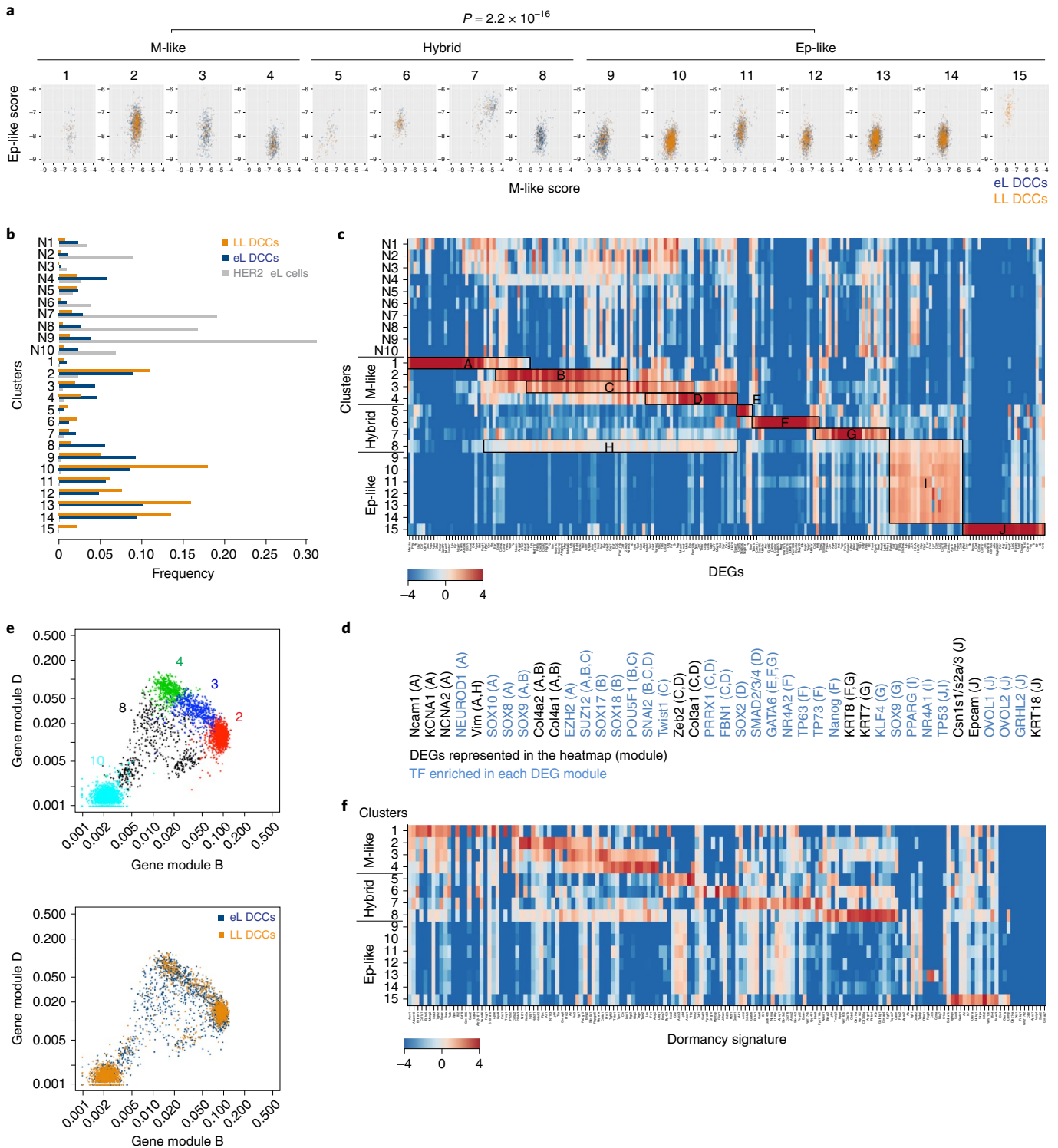
ZFP281 is associated with M-like states in ELs and early DCCs. MMTV-HER2 EL cells do not form tumors but disseminate efficiently and persist as DCCs in lungs²⁰. We hypothesized that the M-like program found in the DCCs may be transcriptionally encoded in ELs. To test this hypothesis, we performed a TF 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 naive pluripotency³⁵, 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 cells³⁶ and it is upregulated during the naive-to-primed pluripotent state transition³⁵ where partial EMT/epithelial plasticity was postulated to happen^{37,38}. However, it is unclear how *ZFP281* regulates EMT and how it is linked to early breast cancer progression. The second largest node, NR5A2/LRH-1 (Fig. 3a), regulates embryonic stem cell pluripotency³⁹, but its link to EMT in breast cancer is unclear^{40,41}. Among other TFs, RARβ and RARγ, previously linked to dormancy⁴², 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 computationally 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 CDH1 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

Fig. 2 | Early DCCs turn on M-like and pluripotency-like programs that associate with dormancy gene signatures. **a**, Distribution of epithelial (Ep) and mesenchymal (M) scores (gene lists in Supplementary Table 4) in MMTV-HER2 lung DCC clusters after unsupervised clustering on the DEGs using a previously described batch-aware algorithm³¹. Cell clusters were subgrouped as M-like (1–4), hybrid (5–8) and Ep-like (9–15). Dots color-coded by sample origin (early lung (eL) DCCs, blue; late lung (LL) DCCs, orange). Pearson's chi-squared test with Yates' continuity correction. **b**, HER2⁻ lung cells (gray) and HER2⁺ eL (blue) and LL (orange) DCC scRNAseq sample distribution per cluster. Unsupervised clustering on the DEGs was performed using a previously described batch-aware algorithm³¹. **c**, Heatmap of UMI counts of selected genes (gene list in Supplementary Table 4) in MMTV-HER2 HER2⁻ lung cells (gray) and HER2⁺ eL (blue), and LL (orange) DCC scRNAseq. N1–N10 are clusters enriched in HER2⁻ lung cells and excluded in further analysis. Clusters 1–15 have less than 16% than HER2⁻ lung cells, so these clusters were considered cancer cell clusters, but HER2⁻ lung cells were excluded in further analysis. Clusters 1–15 DEG modules identified in boxed letters. **d**, DEGs represented in the heatmap in panel c (black) and TFs (blue) enriched in each gene module (in brackets) predicted using Enrichr^{33,34}. Full gene lists in Supplementary Table 5. **e**, Distribution of gene modules B and D in clusters 2, 3, 4, 8 and 9. Dots represent single cells color-coded by cluster (top) and sample origin (bottom). Plot with all cells and clusters in Extended Data Fig. 2e. **f**, Heatmap of UMI counts of genes linked to dormancy phenotypes across different studies (gene list in Supplementary Table 4) in MMTV-HER2 eL and LL DCCs scRNAseq after unsupervised clustering on the DEGs using a previously described batch-aware algorithm³¹. See also Extended Data Figs. 2 and 3.

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 Ecad^{low} (less intense membrane staining) and Twist1^{high} (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 (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



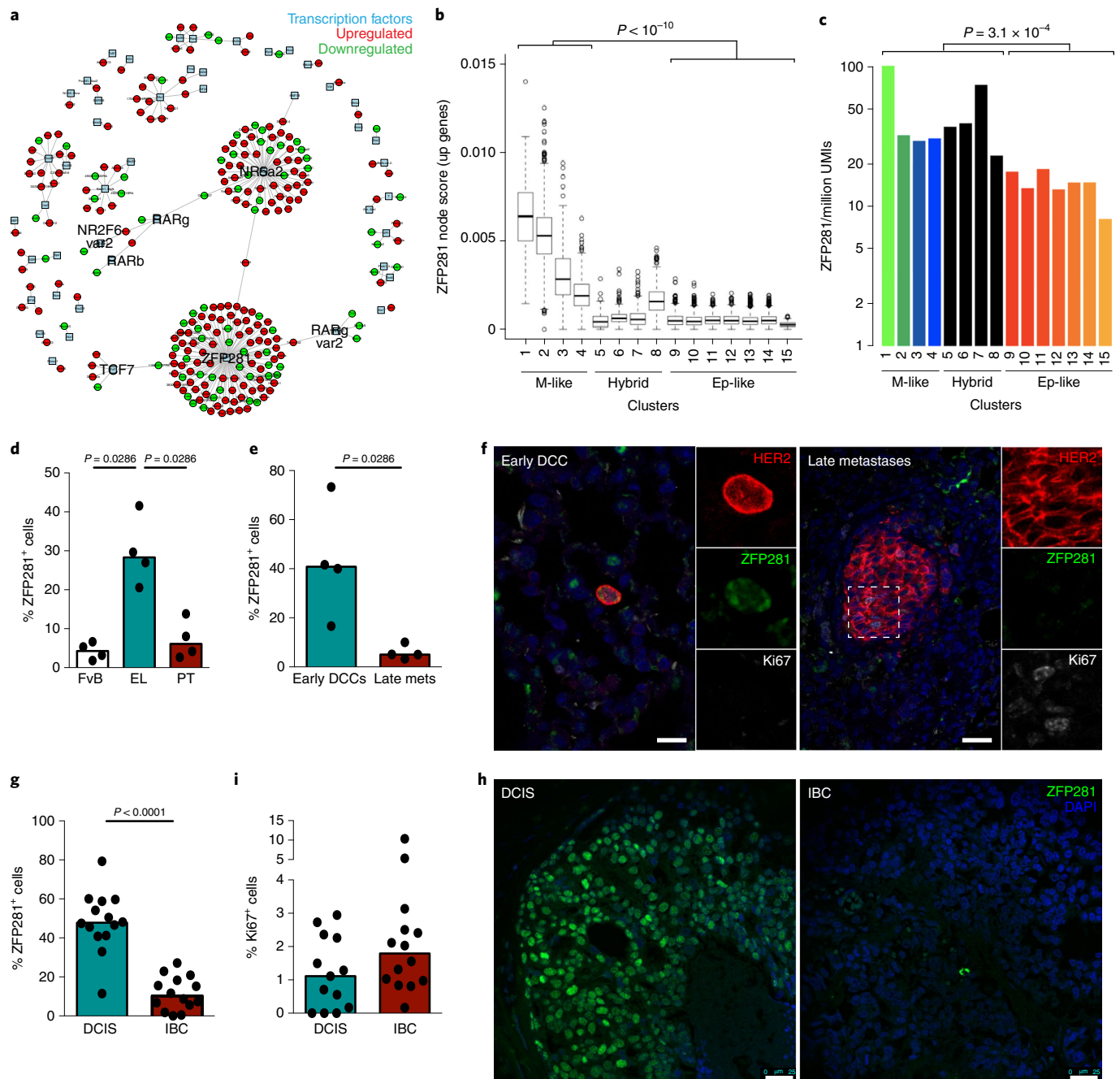


Fig. 3 | Identification of the TF ZFP281 as upregulated in ELs and early DCCs. a, TF (squares and large fonts) network analysis derived from the RNAseq DEGs from MMTV-HER2 EL and PT spheres. Red circles, upregulated genes; green circles, downregulated genes; blue, TFs. Lines connect the TF at the center of the node to target genes, indicating that the connected genes have predicted TF binding elements (validated or predicted) in the promoter regions (-500 to $+2,500$ bp from the transcription start site (TSS)). Full gene list is shown in Supplementary Table 7. **b**, Distribution of ZFP281 predicted target scores, summarizing the UMI fraction of ZFP281 predicted targets (ZFP281 node in Fig. 3a and Supplementary Table 7), in all DCC clusters analyzed by scRNAseq (Fig. 2). Two-tailed Wilcoxon signed-rank test, $P < 10^{-10}$. **c**, Distribution of ZFP281 expression in all DCC clusters analyzed by scRNAseq (Fig. 2). Two-tailed Wilcoxon signed-rank test, $P = 3.1 \times 10^{-4}$. **d**, ZFP281 protein expression in HER2⁺ cells from FvB mammary gland and MMTV-HER2 EL and PT tissues. Graph shows $n = 4$ mice per group, median and two-tailed Mann-Whitney test. Representative pictures are shown in Extended Data Fig. 4b. **e, f**, ZFP281 (green) and Ki67 (gray) protein expression in HER2⁺ cells from MMTV-HER2 (HER2, red) early lung DCCs and late metastasis (mets). Scale bars, 20 μ m. Graph shows $n = 4$ mice per group, median and two-tailed Mann-Whitney test. **g, h**, ZFP281 (green) protein detection in human DCIS and IBC samples. Scale bars, 25 μ m. Graph shows $n = 14$ patients per group, median and two-tailed Mann-Whitney test. **i**, Frequency of Ki67⁺ cells in human DCIS and IBC samples. Graph shows $n = 14$ patients per group, median and nonsignificant two-tailed Mann-Whitney test. See also Extended Data Fig. 4.

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.

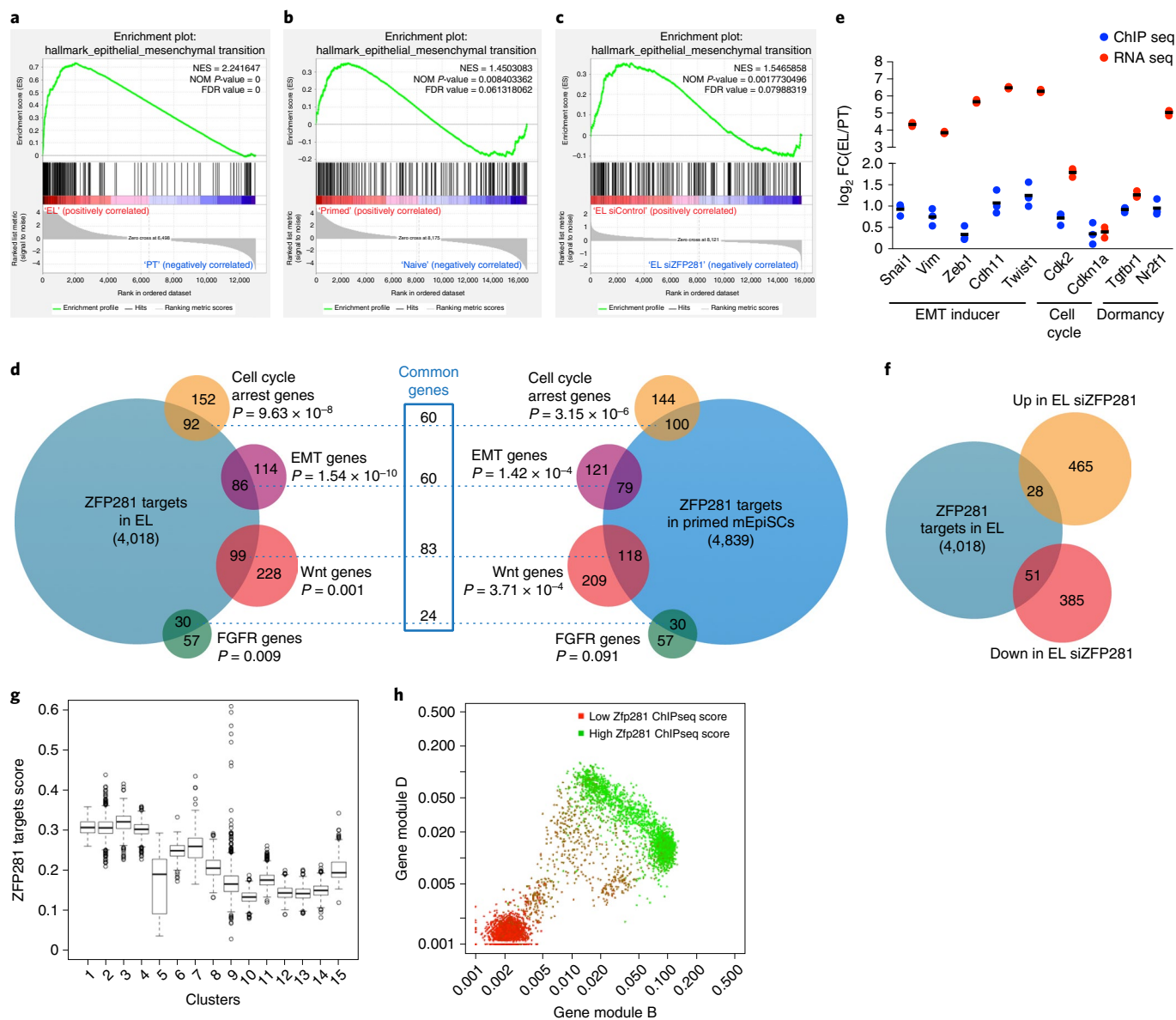


Fig. 4 | ZFP281 regulates EMT, Wnt, FGF and cell cycle programs in both primed mEpiSCs and early DCCs. **a–c**, GSEA^{28,29} of EMT gene expression in MMTV-HER2 EL versus PT (**a**, RNAseq from Fig. 1a), primed versus naive mouse pluripotent stem cells (**b**, RNAseq data: GSE81044 from Fidalgo et al.³⁵) and MMTV-HER2 EL siControl versus siZFP281 (**c**). **d**, Venn diagrams for ZFP281 targets identified by ChIPseq from MMTV-HER2 EL cells (gene list in Supplementary Table 8) and primed mEpiSCs (ChIPseq data: GSE93042 from Huang et al.⁶⁸) and cell cycle arrest, EMT, Wnt and FGF signaling genes. Comparisons and statistics were done in pairs. Overlapped genes between these paired comparisons were rare so for graphical simplification the four comparisons are displayed together. Common genes (central blue box) correspond to genes from each category that are targeted by ZFP281 both in EL and mEpiSCs. Gene lists are shown in Supplementary Table 9. **e**, ChIPseq and RNAseq fold change in MMTV-HER2 EL over PT cells. Graph shows $n = 3$ per condition and median. All genes were upregulated in EL cells; *Snai1*, *Vim*, *Zeb1*, *Cdh11*, *Cdk2*, *Tgfb1* and *Nr2f1* were bound by ZFP281 exclusively in EL cells, whereas *Twist1* and *Cdkn1a* were bound in both EL and PT cells. **f**, Venn diagrams for ZFP281 targets identified by ChIPseq from MMTV-HER2 EL over PT cells and RNAseq from MMTV-HER2 EL siZFP281 over siControl cells. Gene lists are shown in Supplementary Table 12. **g**, Distribution of ZFP281 target (ChIP data) scores, summarizing the averaged expression of ZFP281 targets, in MMTV-HER2 eL (early lungs,) and LL (late lungs) DCCs (scRNAseq, Fig. 2). **h**, Distribution of gene modules B and D in all DCC clusters. Dots represent single cells color-coded by ZFP281 target (ChIP data) scores (low (red) to high (green)). Plot with all cells and clusters in Extended Data Fig. 5i,j. See also Extended Data Fig. 5.

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/naive stem cells were enriched in hallmarks of EMT (Fig. 4a,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 development⁴³.

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 *Snail*, *Vim*, *Zeb1* (EMT inducers²⁷), *Cdk2*, *Cdkn1a* (cell cycle related⁴⁴) and *Tgfb1* and *Nr2f1* (dormancy-associated genes⁴⁵) (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.²⁰ 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 that used in Fig. 4c,f). ZFP281 downregulation in EL spheres leads to a transition from an M-like to Ep-like phenotype, resembling PT spheres (Fig. 5a). FACS analysis showed that upon ZFP281 downregulation, a population of M-like EL cells gain EpCAM expression but do not downregulate Eng expression, leading to an increase in cells with a hybrid phenotype (EpCAM⁺/Eng⁺, dark blue) (Fig. 5d,e). ZFP281 downregulation did not affect the frequency of spheres (Fig. 5b), but it increased the sphere size (number of cells per sphere) (Fig. 5c), supporting a switch to enhanced proliferation upon sphere formation. In Matrigel, the number of invasive acini is significantly lower in the EL shZFP281+DOX condition (Fig. 5f,g). Similar results were observed with an short interfering RNA targeting mouse ZFP281 (Extended Data Fig. 6h,i, targeting sequence used in Fig. 4c,f). Corroborating this partial MET, we observed a decrease in Twist1, Eng and CDH11 (mesenchymal markers) upon

ZFP281 downregulation (Extended Data Fig. 4a, third column). ZFP281 is also downregulated in PTs from the MMTV-PyMT model (Extended Data Fig. 4c), but knockdown of ZFP281 in MMTV-PyMT EL spheres did not enhance proliferation (Extended Data Fig. 6j–l) (see Discussion). Overexpression of ZFP281 in MMTV-HER2 PT spheres (PT ZFP281-OE, Extended Data Fig. 4a, fourth column) induced an invasive M-like phenotype (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 previously²⁰, 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 micrometastasis 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. 7f), 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. 6f), 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

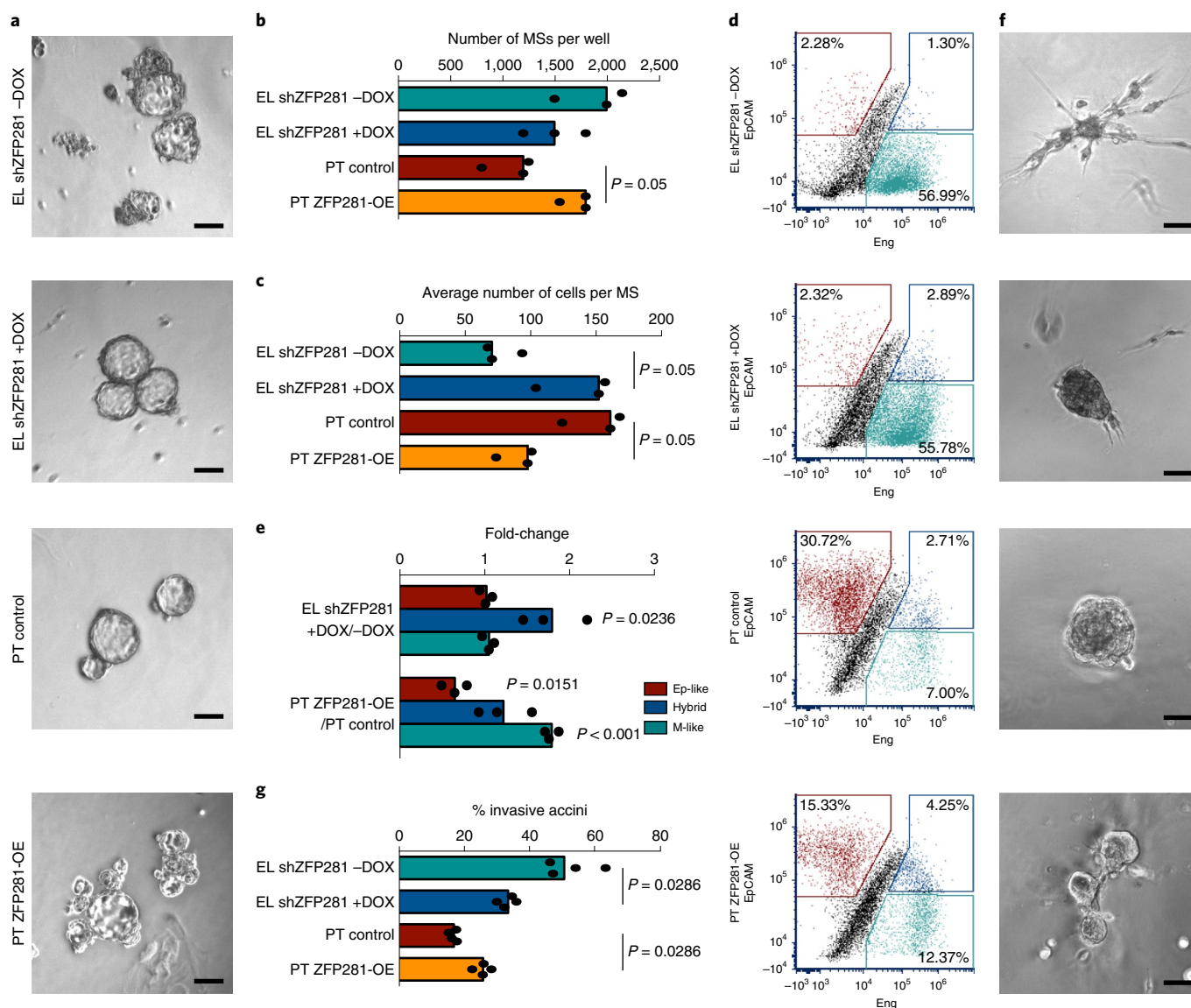


Fig. 5 | ZFP281 induces an M-like invasive and slow-cycling phenotype in vitro. **a**, Representative images of 7-day MMTV-HER2 EL shZFP281±DOX, PT control and PT ZFP281-overexpressed (OE) spheres. Scale bars, 50 μm. **b**, Quantification of mammosphere (MS) frequency of MMTV-HER2 EL shZFP281±DOX, PT 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.

difference in the number of lung single DCCs was found, 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 mEpiSCs 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 ZFP281^{high}, 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.

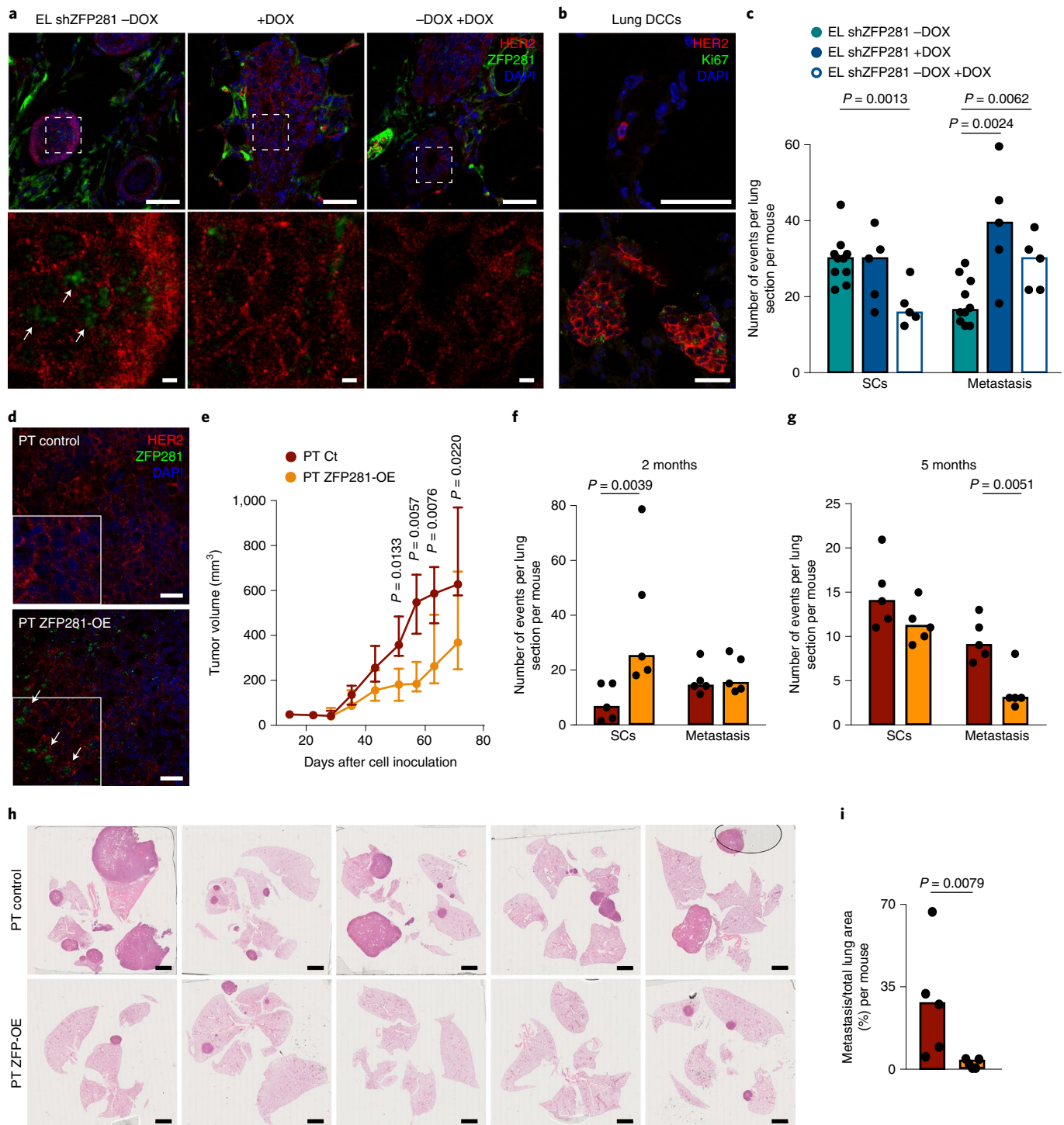


Fig. 6 | ZFP281 favors dissemination but serves as a barrier to metastasis initiation by maintaining an M-like dormancy program in early DCCs in vivo. **a**, Representative images of HER2 (red) and ZFP281 (green) expression in MFPs of mice injected with MMTV-HER2 EL shZFP281 spheres after 5 months. Mice 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 μ m (top row) and 50 μ m (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 μ m. **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 μ m. Quantification in Extended Data Fig. 7f. **e**, Tumor volume growth kinetics of PT control and PT ZFP281-OE spheres injected in the MFP. Tumors were removed at day 71, and mice were killed 5 months after sphere injections (corresponding lungs in panel g). Graph shows $n=10$ tumors per condition, median, interquartile range and multiple t-tests. **f, g**, 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. **h**, H&E images of mice lungs 5 months after sphere injections. Scale bars, 2 mm. **i**, Quantification of lung metastasis burden (images in panel b, one slide 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 marker⁴⁶ upregulated in dormant early DCCs²⁰, 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 FGF2 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^{46–48}, 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^{7,20}. Similarly, early pancreatic DCCs also undergo EMT and persist after seeding the liver²¹. 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^{47,48} may allow early DCCs to maintain the M-like dormant phenotype. Mechanistic analysis supports that FGF2 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 dissemination²⁴, but additional work is needed to determine whether they produce FGF2, or influence the EL or endothelial cells to produce FGF2 and induce ZFP281 expression. FGF2 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 FGF receptors, TWIST1 or other EMT TFs expressed by early DCCs could also awaken these cells from dormancy remains to be tested.

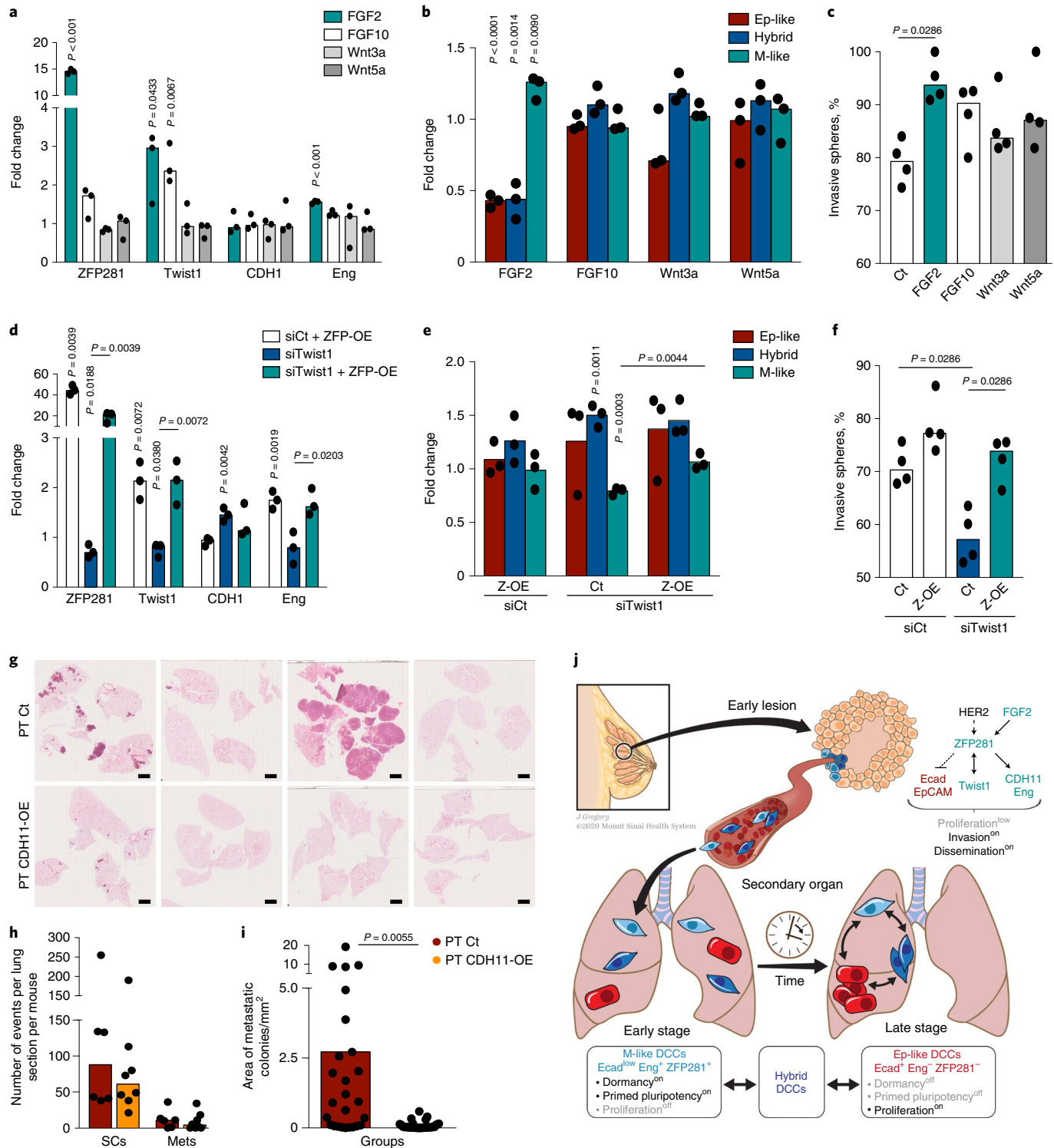
ZFP281 knockdown switched early DCCs to a more hybrid or Ep-like phenotype, which correlates with increased metastatic reactivation. Consistently, Snail- and Zeb1-driven EMT was previously described to suppress cell cycle progression through repression of cyclin D1 and D2 (refs. 49,50). In contrast, MET was associated with rapid relapse and reduced survival in patients with metastatic castrate-resistant prostate cancer⁵¹. 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 proliferative⁵². We previously reported that the lineage commitment regulators DEC2/BHLHE41 and NR2F1/COUP-TF1 coordinate stem-like and quiescence programs^{42,53,54}. However, those studies were in late-evolution cancer models. Our data functionally map these basal/stem-like and developmental/pluripotency programs to such

Fig. 7 | Role of FGF2, TWIST1 and CDH11 in ZFP281-mediated regulation of early DCC fate. a–c, MMTV-HER2 EL spheres and organoids were treated daily with FGF2, FGF10, Wnt3a or Wnt5a for 7 days. Graphs show $n=3$, median with range and two-tailed Mann-Whitney test. **d–f**, MMTV-HER2 EL spheres were transfected with siControl or siTwist1 at days 1 and 2 and cultured for 5 additional days. Graphs show $n=3$, median with range and two-tailed Mann-Whitney test. **a,d**, mRNA expression of *ZFP281*, *Twist1*, *CDH1* and *Eng* in MMTV-HER2 EL spheres. **b,e**, Fold change of Ep-like (EpCAM⁺Eng⁻), hybrid (EpCAM⁺Eng⁺) and M-like (EpCAM⁻Eng⁺) populations in MMTV-HER2 EL spheres. Graphs show $n=3$, median with range and two-tailed Mann-Whitney test. **c,f**, Quantification of 3D-Matrigel invasive phenotype of MMTV-HER2 EL organoids. Graphs show $n=4$, median with range and two-tailed Mann-Whitney test. **g**, H&E images of mice lungs 5 months after MFP injection of PT control or PT *CDH11*-OE spheres. Scale bars, 2 mm. **h,i**, Quantification of lung metastasis burden, shown as number of single cells (SCs) and metastasis per lung section (**h**) and metastasis area (**i**). Graphs show $n=6$ PT control and 8 PT *CDH11*-OE mice, median and two-tailed Mann-Whitney test. **j**, Model of ZFP281-regulated dissemination and dormancy states in early DCCs. Early upon cancer initiation via HER2 signaling, EL cells activate the primed pluripotency TF ZFP281 (top). M-like and pluripotency-like programs lead EL cells to disseminate and to enter a prolonged dormancy as early DCCs in lungs (lower left, early stage). Over time, putative intrinsic and microenvironmental changes allow the dormant DCCs to disrupt ZFP281 function and adopt an Ep-like phenotype (lower right, late stage), which enables a proliferative state. Importantly, M-like, hybrid and Ep-like DCCs coexist in both early- and late-stage lungs, with predominance of M-like dormant DCCs in early-stage lungs. See also Extended Data Fig. 8. Ct, control.

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 dissemination⁵⁶, when detected in prostate and breast cancer DCCs inform on patient prognosis^{42,57}. 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 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 tumors were used as late stage of cancer progression. MMTV-PyMT mice on C57/BL6 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 Institutional Animal Care and Use Committee allowed limit of 1 cm³ per animal.

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 FACS and cell culture preparations, whole mammary glands, PTs and/or lungs were digested in 0.15% Collagenase 1 A (Sigma, C-9891) 2.5% bovine serum albumin (BSA) at 37°C with agitation for 30 min. Red blood cell lysis buffer (Lonza) was used for 2–5 min, and cells were filtered through a 40- μ m filter, passed through a 25G needle and counted. CD45 depletion (MACS, mouse CD45 MicroBeads) was performed for some experiments following the manufacturer's instructions.

MMTV-HER2 mice sphere injection experiments. A total of 300 EL or 150–300 PT spheres were injected per site into nude mice (BALB/c nu/nu, Charles River) in 100 μ l of a 1:1 PBS-Matrigel solution (Corning, growth factor reduced). Spheres were injected in the two fourth inguinal gland fat pad using a 27-gauge needle. Mice injected with sh-TRIPZ-shZFP281 were given control drinking water (–DOX), water with doxycycline from day 0 (+DOX) or water with doxycycline starting 1 month after sphere injection (–DOX + DOX) until the end of the experiment, 5 months after sphere injection. In the case of mice injected with MMTV-HER2 PT spheres, tumors were removed before reaching 1 cm³, according to IAUCU regulations. Mice were euthanized and organs were collected and processed 2 or 5 months after cancer cell injections. Immunofluorescence was performed, and two sections per mice were used to quantify and characterize single DCCs and metastasis. H&E slides were scanned using a NanoZoomer S60 Digital slide scanner and NDP.view2 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 spheres (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 STAR⁵⁸ with default parameters. Read counts for each transcript was measured using featureCounts⁵⁹. DEGs were determined using DESeq2 (ref. ⁶⁰), and a cutoff of 0.05 on adjusted *P* value (corrected for multiple hypotheses 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 GeneJET RNA Purification Kit (K0732, Thermo Fisher Scientific), followed by mRNA library construction using Universal Plus RNA-Seq with NuQuant (0361-A01, TECAN). Libraries were sequenced on the HiSeq 4000 to obtain 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 BioMart⁶¹ to retrieve corresponding genes names. Finally, read counts were normalized with the trimmed mean of M-values method⁶² for differential expression analysis using edgeR (v3.26.8)⁶³.

RNAseq data were further analyzed using Enrichr^{33,34} and GSEA^{28,29}.

Network analysis. Bioinf2bio did this analysis. The genomic sequence corresponding to the promoter (ranging from 2,500 bp upstream from the TSS until 500 bp after the TSS) of each DEG was extracted using the database UCSC (mm10). Next, we retrieved from JASPAR all available position weight matrices corresponding to all known mouse TFs, and for the identification of the TF binding site (TFBS), we used the TFBSTools package⁶⁴ (<http://bioconductor.org/packages/release/bioc/html/TFBSTools.html>) designed to be a computational framework for TF binding analysis. We screened each DEG promoter sequence for all putative TFBS predicted in each of the retrieved position weight matrices. TFBS were scanned in both strands with a 'min.score.percentage' parameter set to 95%. For the network construction, we have used a matrix built from all the TFBS predicted within the promoters of all DEGs selected. Of note, we detected 581,121 TFBSs with a *P* value <10⁻⁴, so we selected only the TFBS with high scores (>21).

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 (Figs. 2–4) non-cancer lung cells (CD45-HER2⁻) and early and late DCCs (CD45-HER2⁺) were sorted. After sorting, cells were encapsulated using the 10X Chromium 3' v2 (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 and final libraries was performed by CyberGreen qPCR library quantification assay (KAPA). Samples were sequenced on an Illumina NextSeq 550 using the 75-cycle kit to a depth of 100 million reads per library. For single-cell clustering, single-cell datasets (first and second experiments) were clustered separately using an unsupervised clustering algorithm previously described³¹. In this expectation–maximization-like algorithm, parameters for multinomial cell-type specific gene-expression models are learned together with a parameter for the fraction of background noise that is associated with cells in each sample. Mitochondrial genes and Malat1 were excluded from the clustering process. The clustering parameters were chosen to accommodate the different UMI and cell counts between the experiments. In the first experiment, the minimum number of UMIs per cell was 300, the number of clusters *k* was 12 and (P₁,P₂) = (30th,60th) percentiles. In the second experiment, the minimum number of UMIs per cell was 800, *k* was set to 20 and (P₁,P₂) = (0th,20th) percentiles. Public GEO records were GSE165456 (first batch) and GSE165459 (second batch). For gene modules and gene scores, gene modules were based on a gene-covariance analysis as was applied previously³¹. Briefly, cells were down-sampled and variable genes were selected based on the variance to mean ratio. Gene-to-gene correlations were estimated per sample and were averaged following z-transformation. The averaged correlation matrices were hierarchically clustered into gene 'modules'. Given a gene list, we calculated its score per cell by summing up the UMIs of the genes in this cell and divided the sum by the total sum of UMIs in the cell. The score therefore equals to the fraction of UMIs associated with the genes in the list.

ChIPseq. ChIP was performed using EZ ChIP protocol (Millipore) with Zfp281 antibody (ab101318, Abcam) for EL and PT samples (after 7 days in cultures). High-throughput sequencing was then used to get the ChIPseq data. ChIPseq 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. ChIPseq peaks were determined by the MACS program (v2.1.2) using input ChIPseq as the control data, and all other parameters followed the default setting. Binding difference around the transcription start sites (–5 kb, +5 kb) between the EL and PT samples are analyzed using the DiffBind (v2.1.6). ChIPseq data were compared with RNA-seq data analysis after RNA-seq 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 (vM19) to count reads for each gene using parameters -r name -f bam, and we used BioMart⁶¹ to retrieve corresponding genes names. Finally, read counts were normalized with the trimmed mean of M-values method⁶² for differential expression analysis using edgeR (v3.26.8)⁶³. Public RNAseq data were downloaded (Key Resource Table) and aligned to mm10, followed with the same processing setting. Public GEO record is GSE165444. Cell cycle arrest, EMT, FGF signaling and Wnt signaling gene sets were downloaded from MsigDB (<https://www.gsea-msigdb.org/gsea/msigdb/index.jsp>) with systematic names M1134, M5930, M1090 and M7847, respectively. All processed and index sorted BAM files of high-throughput sequencing data were converted to TDF files using count command of igvtools, followed by visualization using IGV software^{65,66}.

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). 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, 5×10^6 cells were seeded in six-well ultralow-adhesion plates in 1 ml mammosphere media (DMEM/F12 (Gibco, 11320-082), 1:50 B27 (Invitrogen, 17504-044), 500 ng ml⁻¹ hydrocortisone (Lonza CC-403), 40 µg ml⁻¹ Insulin (Gibco 12585-014), 20 ng ml⁻¹ EGF (Peprotech AF-100-15-A), 100 U ml⁻¹ penicillin and 100 µg ml⁻¹ 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 measurement of sphere size.

EL cells were transfected 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⁻¹ DOX. EL or PT cells were transfected with ZFP281-OE plasmid (pB-3XFL-ZFP281), control vector (Sino Biological, pCMV3-C-GFPspark-CV), CDH11-OE (Sino Biological, MG-51164-ACG), siZFP281 (OriGene, SR421015), siCDH11 (Ambion, 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.

For organoid cultures, cells were seeded in eight-well chamber slides coated with 50 µl Matrigel (Corning, growth factor reduced) per slide in 400 µl assay medium (DMEM/F12, 5% horse serum, 500 ng ml⁻¹ hydrocortisone (Lonza, CC-403), 40 µg ml⁻¹ insulin (Gibco, 12585-014), 100 U ml⁻¹ penicillin and 100 µg ml⁻¹ streptomycin (Corning) and 2% Matrigel (Corning, growth factor reduced). Four pictures of random fields per well were analyzed to quantify the percentage of invasive structures. All in vitro experiments were performed and analyzed using four wells per condition (technical replicates) and at least three independent experiments (biological replicates).

Immunofluorescence. Tissue slides (see Animal experiments section) were dehydrated, followed by antigen retrieval 10 mM citrate buffer, pH 6.0 (Na₂H₆O₇). Blocking was done using 0.5% BSA in PBS with 5% normal goat serum (Thermo Fisher Scientific, PCN5000) 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 Å~ immunofluorescence PBS wash buffer (130 mM NaCl, 7 mM Na₂HPO₄, 3.5 mM NaH₂PO₄, 7.7 mM NaNO₃, 0.1 % BSA, 0.2% Triton X-100 and 0.05% Tween-20) containing 5% normal goat serum (Thermo Fisher Scientific, PCN5000) for 1 h. Antibodies and conditions used are summarized in the Reporting summary. Chambers were removed from slides, and wells were fixed and 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-CT 1-Step Power SYBR Green kit (Invitrogen, A25600) and primers from Supplementary Table 13. GAPDH or RPL0 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 quantifications 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.

Reporting summary. 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 (GSE165431, RNAseq of MMTV-Neu EL and PT spheres; GSE165444, 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: 01 September 2022

References

- Klein, K. C. Cancer progression and the invisible phase of metastatic colonization. *Nat. Rev. Cancer* **20**, 681–694 (2020).
- Pantel, K. & Brakenhoff, R. H. Dissecting the metastatic cascade. *Nat. Rev. Cancer* **4**, 448 (2004).
- Valastyan, S. & Weinberg, R. A. Tumor metastasis: molecular insights and evolving paradigms. *Cell* **147**, 275–292 (2011).
- Yates, L. R. et al. Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell* **32**, 169–184 (2017).
- Schardt, J. A. et al. Genomic analysis of single cytokeratin-positive cells from bone marrow reveals early mutational events in breast cancer. *Cancer Cell* **8**, 227–239 (2005).
- 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).
- Hosseini, H. et al. Early dissemination seeds metastasis in breast cancer. *Nature* **540**, 552–558 (2016).
- Ullah, I. et al. Evolutionary history of metastatic breast cancer reveals minimal seeding from axillary lymph nodes. *J. Clin. Invest.* **128**, 1355–1370 (2018).
- Casasent, A. K. et al. Multiclonal invasion in breast tumors identified by topographic single cell sequencing. *Cell* **172**, 205–217 (2018).
- 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).
- Rhim, A. D. et al. Detection of circulating pancreas epithelial cells in patients with pancreatic cystic lesions. *Gastroenterology* **146**, 647–651 (2014).
- Makohon-Moore, A. P. et al. Precancerous neoplastic cells can move through the pancreatic ductal system. *Nature* **561**, 201–205 (2018).
- Muzumdar, M. D. et al. Clonal dynamics following p53 loss of heterozygosity in Kras-driven cancers. *Nat. Commun.* **7**, 12685 (2016).
- Shain, A. H. et al. The genetic evolution of metastatic uveal melanoma. *Nat. Genet.* **51**, 1123–1130 (2019).
- Hu, Z. et al. Quantitative evidence for early metastatic seeding in colorectal cancer. *Nat. Genet.* **51**, 1113–1122 (2019).
- Turajlic, S. et al. Tracking cancer evolution reveals constrained routes to metastases: TRACERx Renal. *Cell* **173**, 581–594 (2018).
- 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).
- 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).
- Wang, D. et al. Multiregion sequencing reveals the genetic heterogeneity and evolutionary history of osteosarcoma and matched pulmonary metastases. *Cancer Res.* **79**, 7–20 (2019).
- Harper, K. L. et al. Mechanism of early dissemination and metastasis in Her2+ mammary cancer. *Nature* **540**, 588–592 (2016).
- Rhim, A. D. et al. EMT and dissemination precede pancreatic tumor formation. *Cell* <https://doi.org/10.1016/j.cell.2011.11.025> (2012).
- 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).

23. Eyles, J. et al. Tumor cells disseminate early, but immunosurveillance limits metastatic outgrowth, in a mouse model of melanoma. *J. Clin. Invest.* **120**, 2030–2039 (2010).
24. Linde, N. et al. Macrophages orchestrate breast cancer early dissemination and metastasis. *Nat. Commun.* **9**, 21 (2018).
25. Guy, C. T. et al. Expression of the neu protooncogene in the mammary epithelium of transgenic mice induces metastatic disease. *Proc. Natl Acad. Sci. USA* **89**, 10578–10582 (1992).
26. Lu, P., Takai, K., Weaver, V. M. & Werb, Z. Extracellular matrix degradation and remodeling in development and disease. *Cold Spring Harb. Perspect. Biol.* **3**, a005058 (2011).
27. Nieto, M. A., Huang, R. Y. Y. J., Jackson, R. A. A. & Thiery, J. P. P. EMT: 2016. *Cell* **166**, 21–45 (2016).
28. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
29. Mootha, V. K. et al. PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
30. Lim, E. et al. Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. *Breast Cancer Res.* **12**, R21 (2010).
31. Martin, J. C. et al. Single-cell analysis of Crohn's disease lesions identifies a pathogenic cellular module associated with resistance to anti-TNF therapy. *Cell* **178**, 1493–1508 (2019).
32. Barbara, N. P., Wrana, J. L. & Letarte, M. Endoglin is an accessory protein that interacts with the signaling receptor complex of multiple members of the transforming growth factor- β superfamily. *J. Biol. Chem.* <https://doi.org/10.1074/jbc.274.2.584> (1999).
33. Chen, E. Y. et al. Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
34. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
35. Fidalgo, M. et al. Zfp281 Coordinates opposing functions of Tet1 and Tet2 in pluripotent states. *Cell Stem Cell* **19**, 355–369 (2016).
36. Hahn, S., Jackstadt, R., Siemens, H., Hünten, S. & Hermeking, H. SNAIL and miR-34a feed-forward regulation of ZNF281/ZBP99 promotes epithelial-mesenchymal transition. *EMBO J.* **32**, 3079–3095 (2013).
37. Micalizzi, D. S. et al. Epithelial-mesenchymal transition in cancer: parallels between normal development and tumor progression. *J. Mammary Gland Biol. Neoplasia* **15**, 117–134 (2010).
38. Kim, D. et al. Epithelial mesenchymal transition in embryonic development, tissue repair and cancer: a comprehensive overview. *J. Clin. Med.* **7**, 1 (2017).
39. Gu, P. et al. Orphan nuclear receptor LRH-1 is required to maintain Oct4 expression at the epiblast stage of embryonic development. *Mol. Cell. Biol.* **25**, 3492–3505 (2005).
40. Liu, L. et al. Nr5a2 promotes tumor growth and metastasis of gastric cancer AGS cells by Wnt/beta-catenin signaling. *Oncotargets Ther.* **12**, 2891–2902 (2019).
41. Luo, Z. et al. Effect of NR5A2 inhibition on pancreatic cancer stem cell (CSC) properties and epithelial-mesenchymal transition (EMT) markers. *Mol. Carcinog.* **56**, 1438–1448 (2017).
42. Sosa, M. S. et al. NR2F1 controls tumour cell dormancy via SOX9- and RAR β -driven quiescence programmes. *Nat. Commun.* **6**, 1–14 (2015).
43. Aclouque, H. et al. Epithelial-mesenchymal transitions: the importance of changing cell state in development and disease. *J. Clin. Invest.* **119**, 1438–1439 (2009).
44. Johnson, D. G. & Walker, C. L. Cyclins and cell cycle checkpoints. *Annu. Rev. Pharmacol. Toxicol.* **39**, 295–312 (1999).
45. Risson, E., Nobre, A. R., Maguer-Satta, V. & Aguirre-Ghiso, J. A. The current paradigm and challenges ahead for the dormancy of disseminated tumor cells. *Nat. Cancer* **1**, 672–680 (2020).
46. Zeisberg, M. & Neilson, E. G. Biomarkers for epithelial-mesenchymal transitions. *J. Clin. Invest.* **119**, 1429–1437 (2009).
47. Johnson, C. L. et al. Impaired macrophage trafficking and increased helper T-cell recruitment with loss of cadherin-11 in atherosclerotic immune response. *Am. J. Physiol. Heart Circ. Physiol.* **321**, H756–H769 (2021).
48. Schneider, D. J. et al. Cadherin-11 contributes to pulmonary fibrosis: potential role in TGF- β production and epithelial to mesenchymal transition. *FASEB J.* **26**, 503–512 (2012).
49. Vega, S. et al. Snail blocks the cell cycle and confers resistance to cell death. *Genes Dev.* **18**, 1131–1143 (2004).
50. Mejlvang, J. et al. Direct repression of cyclin D1 by SIP1 attenuates cell cycle progression in cells undergoing an epithelial mesenchymal transition. *Mol. Biol. Cell* **18**, 4615–4624 (2007).
51. Stylianou, N. et al. A molecular portrait of epithelial-mesenchymal plasticity in prostate cancer associated with clinical outcome. *Oncogene* **38**, 913–934 (2019).
52. Lawson, D. A. et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature* **526**, 131–135 (2015).
53. Adam, A. P. et al. Computational identification of a p38 SAPK-regulated transcription factor network required for tumor cell quiescence. *Cancer Res.* **69**, 5664–5672 (2009).
54. Bragado, P. et al. TGF- β 2 dictates disseminated tumour cell fate in target organs through TGF- β -RIII and p38 α / β signalling. *Nat. Cell Biol.* **15**, 1351–1361 (2013).
55. Laughney, A. M. et al. Regenerative lineages and immune-mediated pruning in lung cancer metastasis. *Nat. Med.* **26**, 259–269 (2020).
56. Rodriguez-Tirado, C. et al. NR2F1 is a barrier to dissemination of early stage breast cancer cells. *Cancer Res.* **82**, 2313–2326 (2022).
57. Borgen, E. et al. NR2F1 stratifies dormant disseminated tumor cells in breast cancer patients. *Breast Cancer Res.* **20**, 120 (2018).
58. Dobin, A. et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
59. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
60. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
61. Durinck, S. et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
62. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
63. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
64. Tan, G. & Lenhard, B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32**, 1555–1556 (2016).
65. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant review with the integrative genomics viewer. *Cancer Res.* **77**, e31–e34 (2017).
66. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings Bioinform.* **14**, 178–192 (2013).
67. Fidalgo, M. et al. Zfp281 mediates Nanog autorepression through recruitment of the NuRD complex and inhibits somatic cell reprogramming. *Proc. Natl Acad. Sci. USA* **109**, 16202–16207 (2012).
68. Huang, X. et al. Zfp281 is essential for mouse epiblast maturation through transcriptional and epigenetic control of Nodal signaling. *Elife* **6**, e33333 (2017).

Acknowledgements

We thank the Aguirre-Ghiso lab for helpful discussions and thank the expertise and assistance of the Dean's Flow Cytometry CoRE and Microscope CoRE, Icahn School of Medicine at Mount Sinai. We thank J.K. Gregory, medical illustrator at Mount Sinai, for the graphical abstract. This work was supported by the National Institutes of Health/National Cancer Institute (CA109182, CA216248, CA218024 and CA196521) and the Samuel Waxman Cancer Research Foundation Tumor Dormancy Program. A.R.N. was funded by the Portuguese Foundation for Science and Technology (SFRH/BD/100380/2014). E.D. was funded by the National Institutes of Health/National Cancer Institute (T32 CA078207). L.W. was funded by the Federal Ministry of Education and Research (DFG project number 431474090). Research in the laboratory of J.W. was supported by grants from NYSTEM (C32569GG and C32583GG) and the National Institutes of Health (R01GM129157, R01HD095938, R01HD097268 and R01HL146664).

Author contributions

A.R.N. designed, planned and conducted experiments, analyzed data and wrote the manuscript; E.D., J.Y., X.H., L.W., E.R., P.R., M.L.A. and W.Z. conducted experiments; J.Y. and E.K. did the bioinformatics analysis; J.A.S. did the bioinformatics analysis in human datasets; CC provided expertise on analysis of human datasets; J.W. provided necessary reagents and developmental biology expertise; and J.A.A.-G. designed experiments, analyzed data and wrote the manuscript.

Competing interests

J.A.A.-G. is a scientific co-founder of scientific advisory board member and equity owner in HiberCell and receives financial compensation as a consultant for HiberCell, a Mount Sinai spin-off company focused on the research and development of therapeutics that prevent or delay the recurrence of cancer. The remaining authors declare no competing interests.

Additional information

Extended data are available for this paper at <https://doi.org/10.1038/s43018-022-00424-8>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s43018-022-00424-8>.

Correspondence and requests for materials should be addressed to Ana Rita Nobre or Julio A. Aguirre-Ghiso.

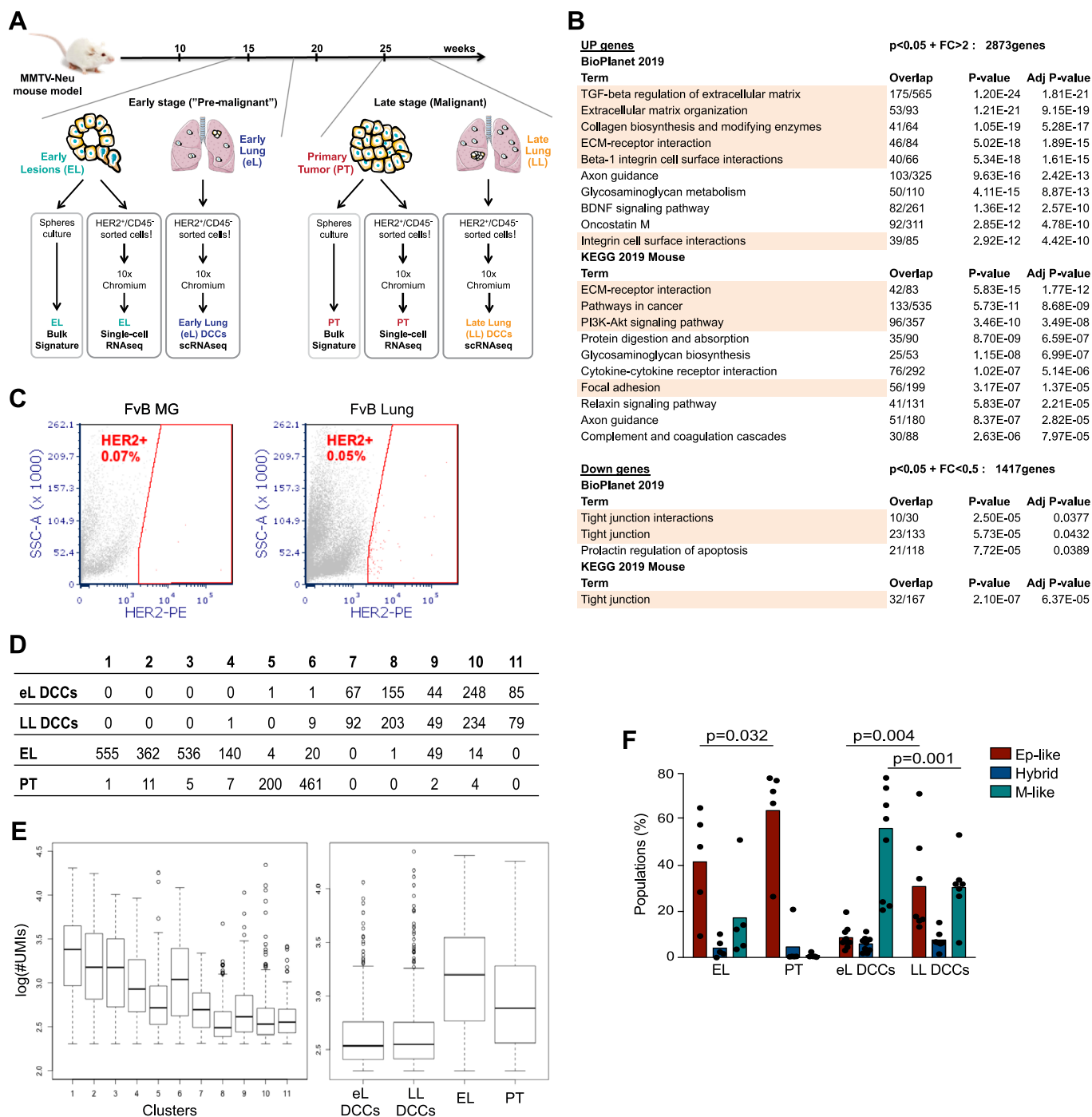
Peer review information *Nature Cancer* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

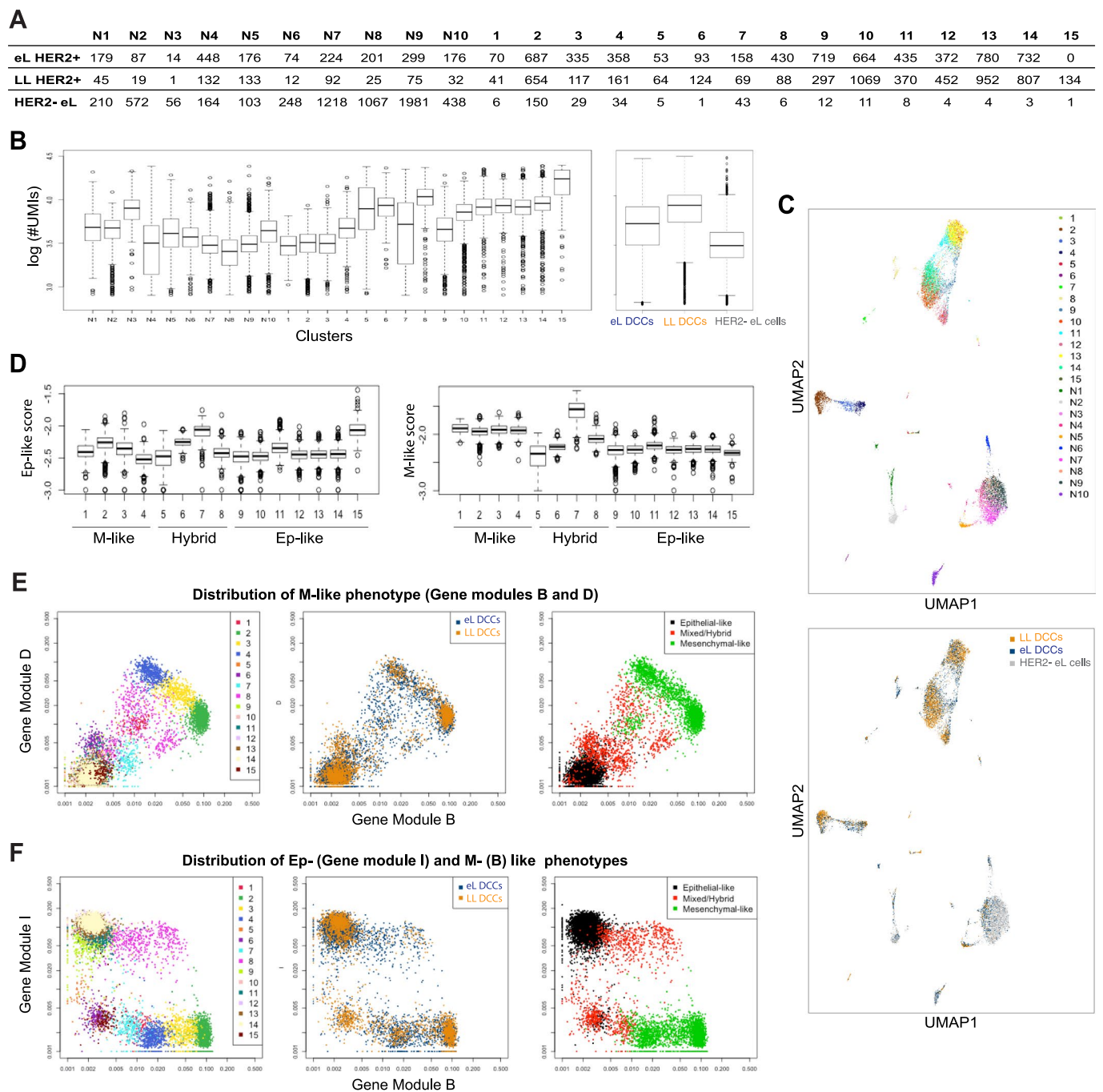
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

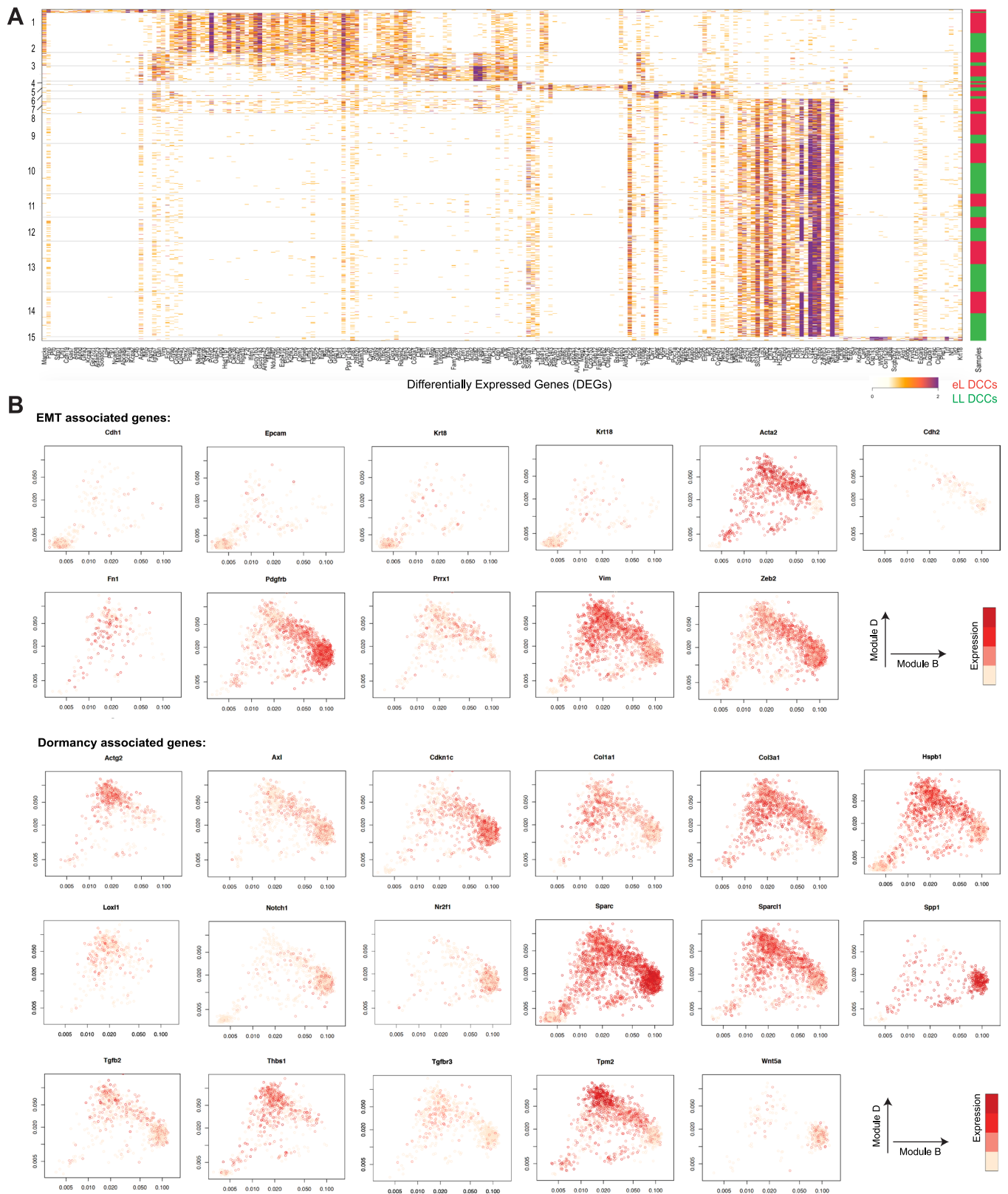
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022



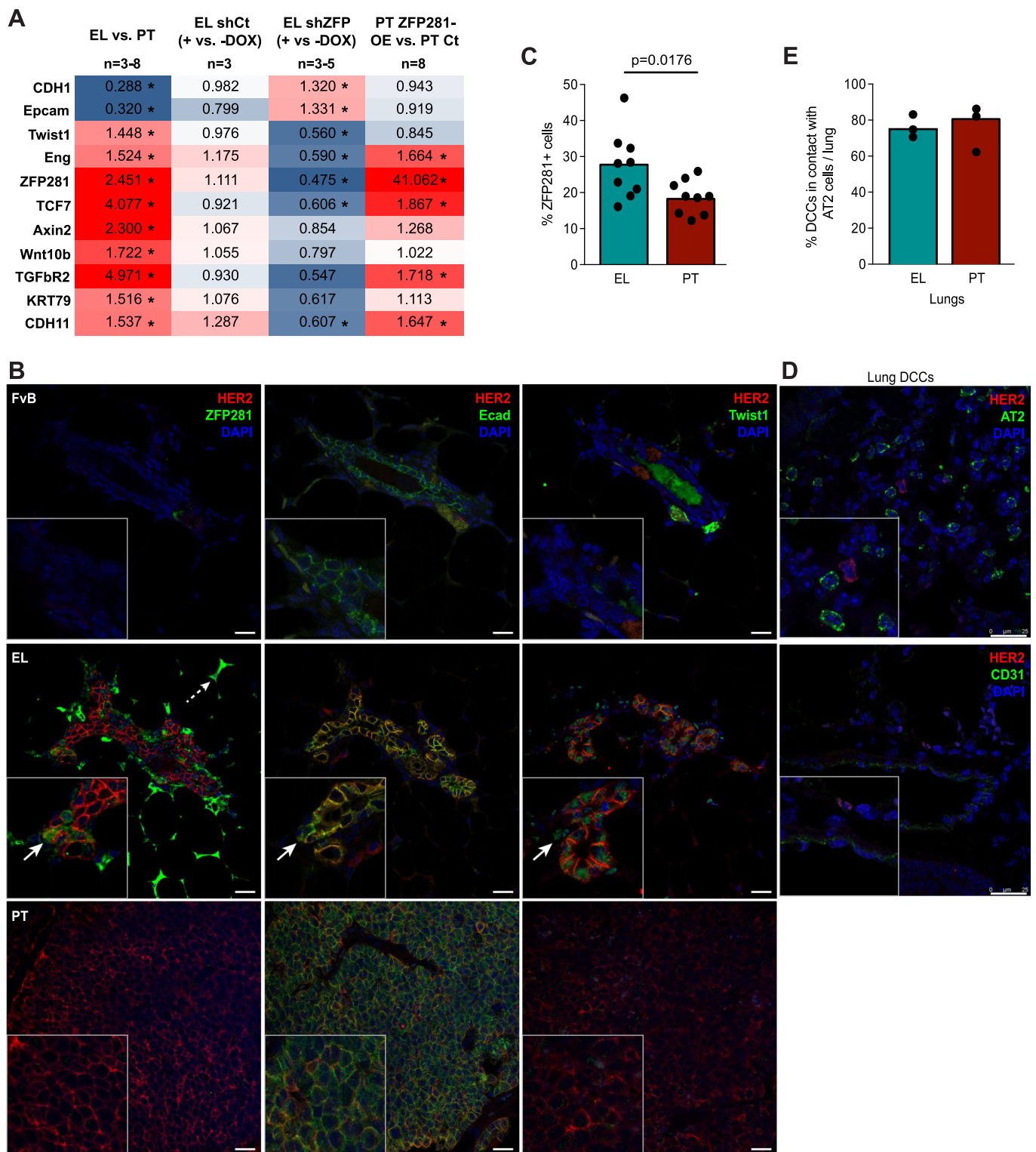
Extended Data Fig. 1 | Model description, Enrichr analysis, FACS controls and MMTV-HER2 scRNAseq data distribution. (a) Experimental design of MMTV-HER2 bulk and single-cell RNA sequencing experiments. (b) Enrichr analysis^{30,31} of differentially expressed genes (DEG) in MMTV-HER2 early lesion (EL) 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.



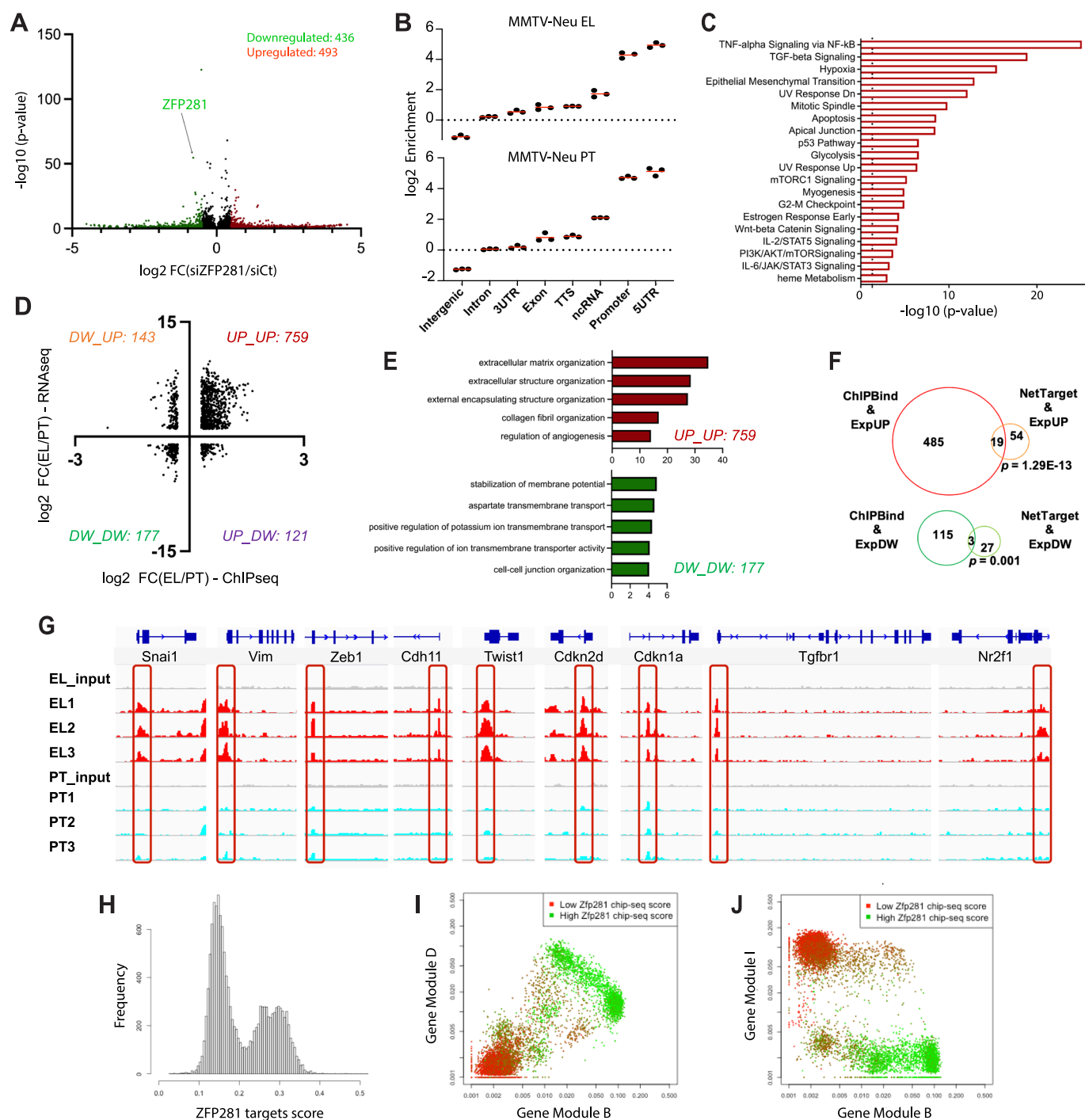
Extended Data Fig. 2 | MMTV-HER2 scRNAseq data distribution of phenotypes across clusters. (a) Number of cells per cluster analyzed in the single-cell RNAseq of HER2, HER2⁺ eL (early lungs) and LL (late lungs) DCCs (see Fig. 2b, c). (b) Number of UMIs per cluster (left) and per sample (right) analyzed 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.



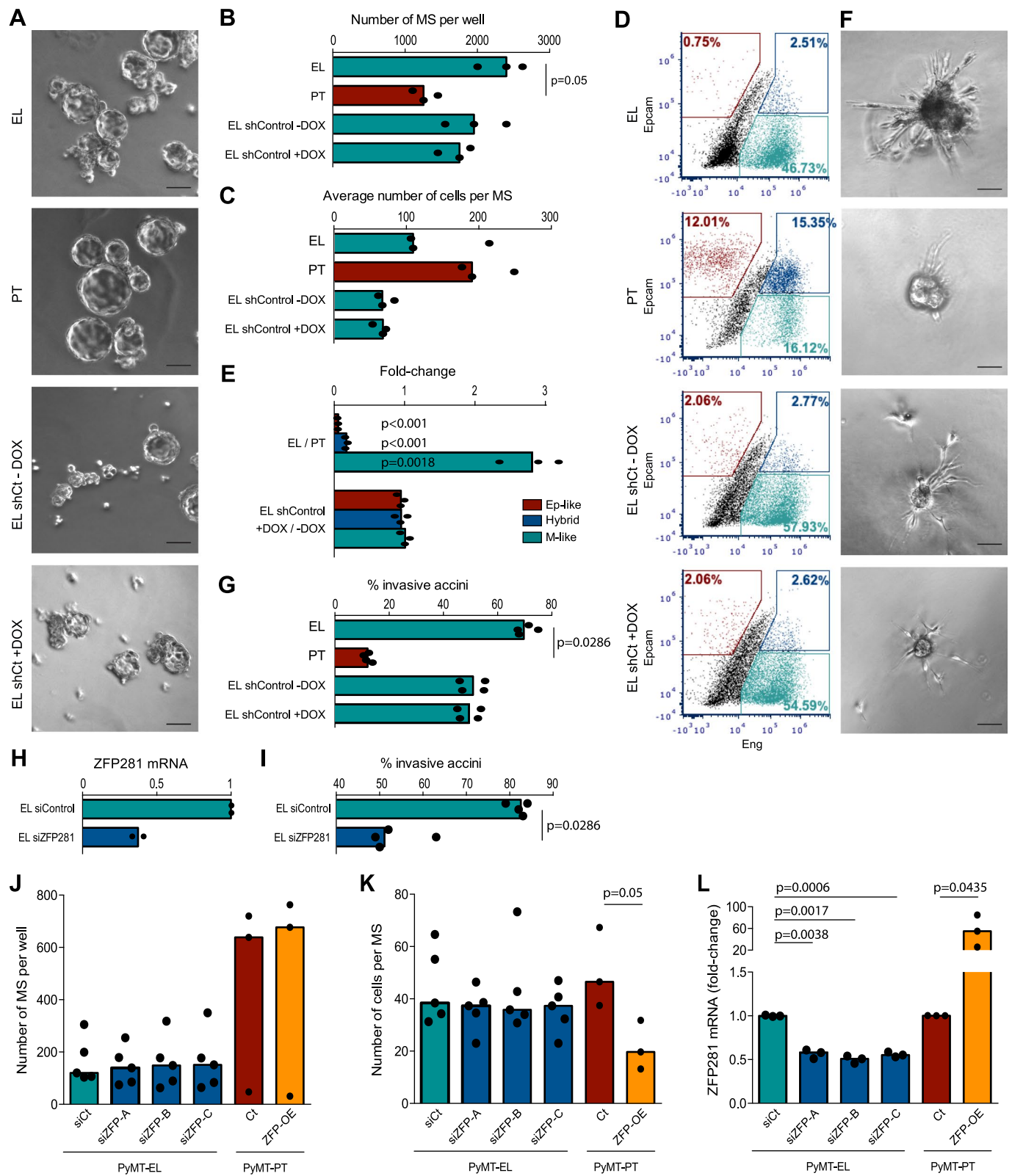
Extended Data Fig. 3 | High resolution plots for scRNASeq clusters and projections on to M-like signatures. (a) Heatmap of UMI counts of selected genes (gene lists in Supplementary Table 4) in MMTV-HER2 eL (early lungs,) and LL (late lungs) DCCs single-cell RNAseq after unsupervised clustering on the 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-value <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⁺Ecad^{low}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.



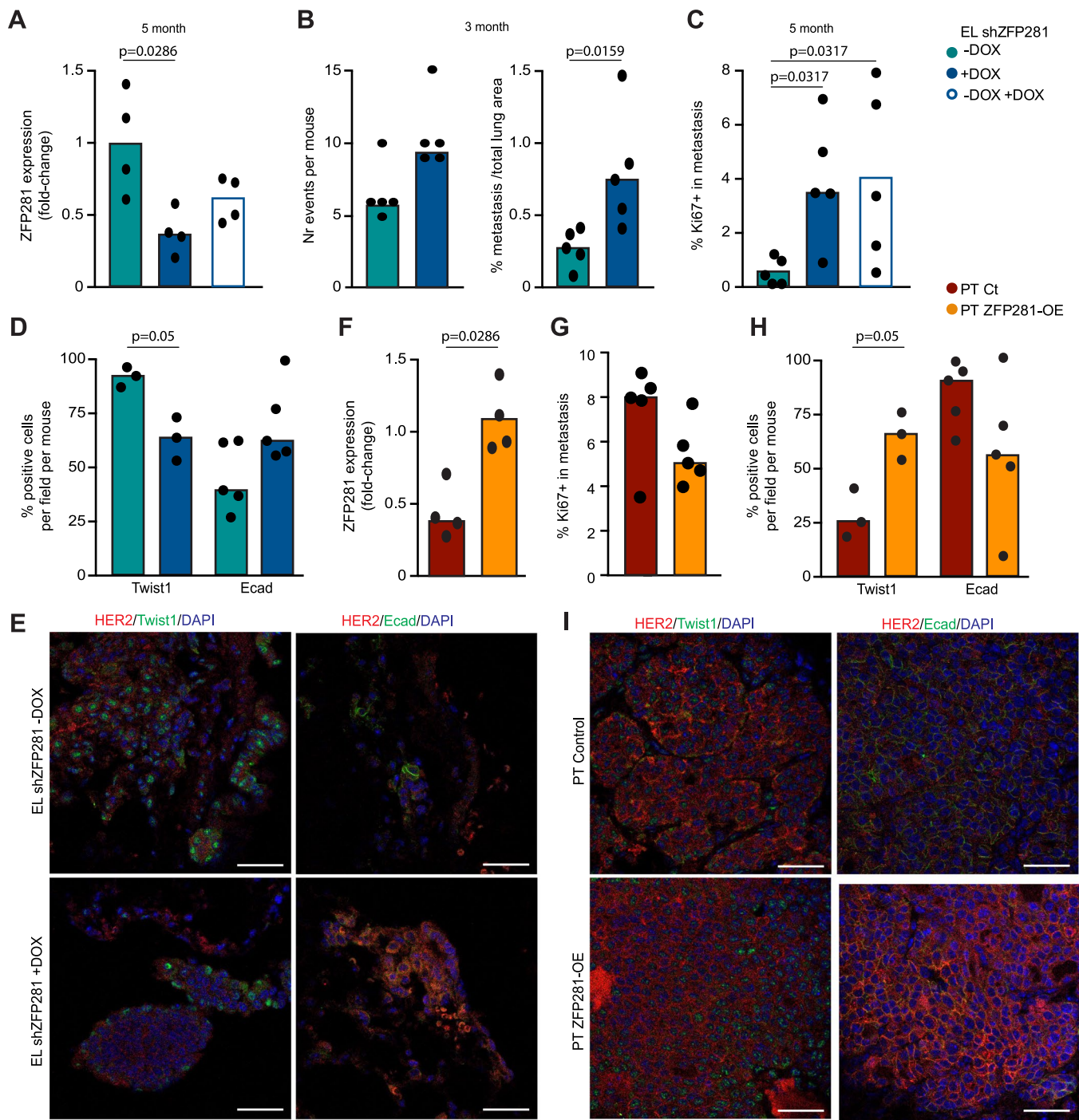
Extended Data Fig. 5 | RNAseq and ChIP-seq analysis for ZFP281 targets and projection of ZFP281 binding score on to M-, H- and Ep-like clusters. (a) Volcano plot from RNAseq data on MMTV-HER2 siZFP281 cells. 436 downregulated genes (green), p -value <0.05 & $\log_2FC > -0.5$; 493 upregulated genes (red), p -value <0.05 & $\log_2FC > 0.5$. 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, Twist1; Cell cycle associated genes: Cdkn2d, Cdkn1a; Dormancy-associated genes: Tgfb1, 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).



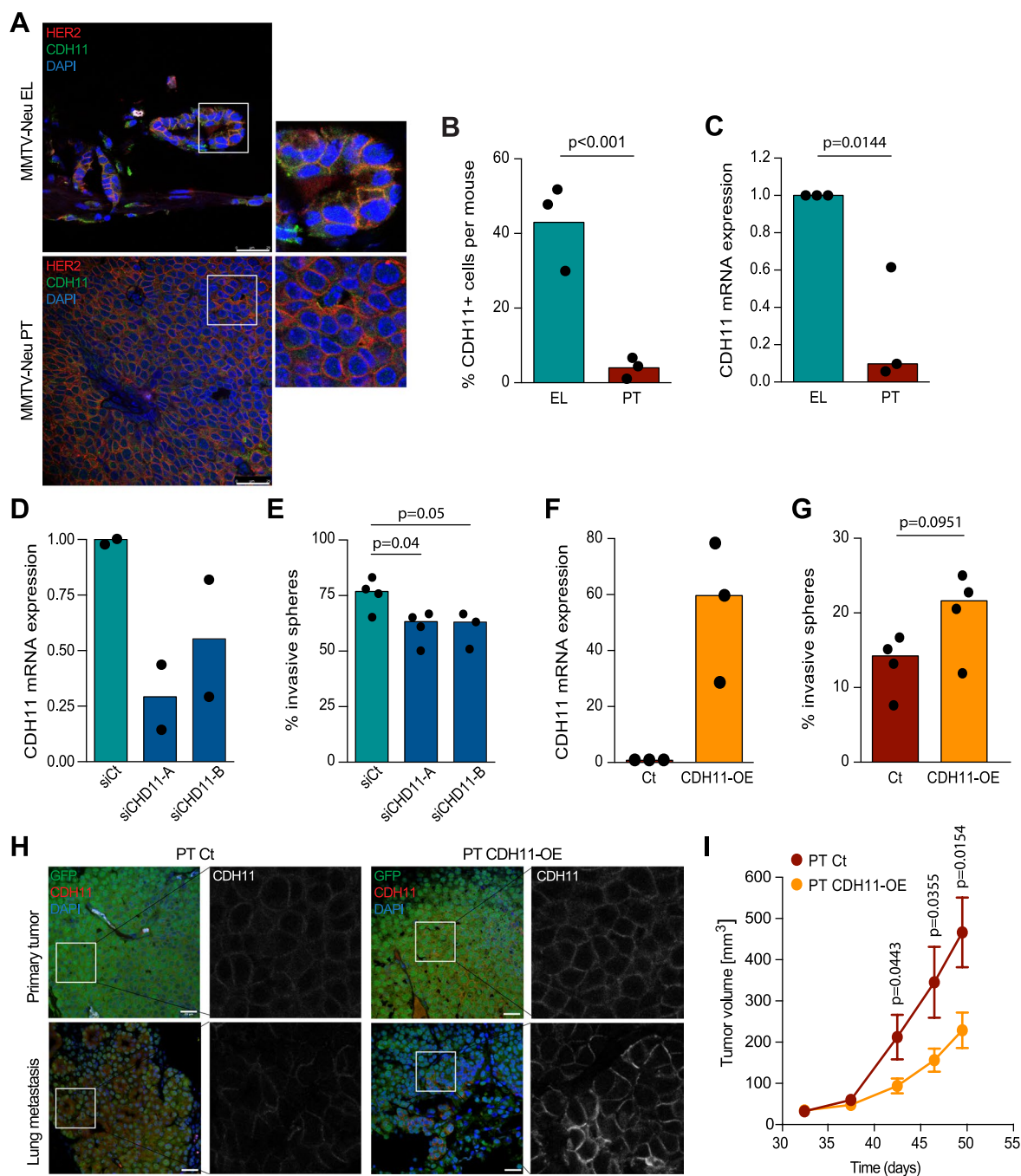
Extended Data Fig. 6 | See next page for caption.

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.