ARTICLES
https://doi.org/10.1038/s41588-021-00911-1

nature genetics

Check for updates

A single-cell and spatially resolved atlas of human breast cancers

Sunny Z. Wu[1,2,26], Ghamdan Al-Eryani[1,2,26], Daniel Lee Roden[1,2,26], Simon Junankar[1,2], Kate Harvey[1], Alma Andersson[3], Aatish Thennavan[4], Chenfei Wang[5], James R. Torpy[1,2], Nenad Bartonicek[1,2], Taopeng Wang[1,2], Ludvig Larsson[3], Dominik Kaczorowski[6], Neil I. Weisenfeld[7], Cedric R. Uytingco[7], Jennifer G. Chew[7], Zachary W. Bent[7], Chia-Ling Chan[6], Vikkitharan Gnanasambandapillai[6], Charles-Antoine Dutertre[8,9], Laurence Gluch[10], Mun N. Hui[1,11], Jane Beith[11], Andrew Parker[2,12], Elizabeth Robbins[13], Davendra Segara[12], Caroline Cooper[14,15], Cindy Mak[16,17], Belinda Chan[16], Sanjay Warrier[16,17], Florent Ginhoux[18,19,20], Ewan Millar[21,22,23], Joseph E. Powell[6,24], Stephen R. Williams[7], X. Shirley Liu[5], Sandra O'Toole[1,13,23,25], Elgene Lim[1,2,12], Joakim Lundeberg[3], Charles M. Perou[4] and Alexander Swarbrick[1,2] ✉

Breast cancers are complex cellular ecosystems where heterotypic interactions play central roles in disease progression and response to therapy. However, our knowledge of their cellular composition and organization is limited. Here we present a single-cell and spatially resolved transcriptomics analysis of human breast cancers. We developed a single-cell method of intrinsic subtype classification (SCSubtype) to reveal recurrent neoplastic cell heterogeneity. Immunophenotyping using cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) provides high-resolution immune profiles, including new PD-L1/PD-L2$^+$ macrophage populations associated with clinical outcome. Mesenchymal cells displayed diverse functions and cell-surface protein expression through differentiation within three major lineages. Stromal-immune niches were spatially organized in tumors, offering insights into antitumor immune regulation. Using single-cell signatures, we deconvoluted large breast cancer cohorts to stratify them into nine clusters, termed 'ecotypes', with unique cellular compositions and clinical outcomes. This study provides a comprehensive transcriptional atlas of the cellular architecture of breast cancer.

Breast cancers are clinically stratified based on the expression of the estrogen receptor (ER), progesterone receptor (PR) and overexpression of human epidermal growth factor receptor 2 (HER2) or amplification of the HER2 gene *ERBB2*. This results in three broad subtypes that correlate with prognosis and define treatment strategies: luminal (ER$^+$, PR$^{+/-}$); HER2$^+$ (HER2$^+$, ER$^{+/-}$, PR$^{+/-}$); and triple negative breast cancer (TNBC; ER$^-$, PR$^-$, HER2$^-$). Breast cancers are also stratified based on bulk transcriptomic profiling using the PAM50 gene signature into five 'intrinsic' molecular subtypes: luminal-like (LumA and LumB); HER2-enriched (HER2E); basal-like; and normal-like. There is an approximate 70–80% concordance between molecular and clinical subtypes[1,2]. While PAM50 has provided important insights into prognosis and treatment[3–6], the functional understanding of these subtypes at cellular resolution is currently limited.

[1]The Kinghorn Cancer Centre and Cancer Research Theme, Garvan Institute of Medical Research, Darlinghurst, New South Wales, Australia. [2]St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Sydney, New South Wales, Australia. [3]Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Solna, Sweden. [4]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [5]Department of Data Science, Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [6]Garvan-Weizmann Centre for Cellular Genomics, Garvan Institute of Medical Research, Sydney, New South Wales, Australia. [7]10x Genomics, Pleasanton, CA, USA. [8]Gustave Roussy Cancer Campus, Villejuif, France. [9]Institut National de la Santé Et de la Recherche Médicale (INSERM) U1015, Equipe Labellisée—Ligue Nationale contre le Cancer, Villejuif, France. [10]The Strathfield Breast Centre, Strathfield, New South Wales, Australia. [11]Chris O'Brien Lifehouse, Camperdown, New South Wales, Australia. [12]St Vincent's Hospital, Darlinghurst, New South Wales, Australia. [13] Department of Tissue Pathology and Diagnostic Pathology, New South Wales Health Pathology, Royal Prince Alfred Hospital, Camperdown, New South Wales, Australia. [14]Pathology Queensland, Princess Alexandra Hospital, Brisbane, Queensland, Australia. [15]Southside Clinical Unit, Faculty of Medicine, University of Queensland, Brisbane, Queensland, Australia. [16]Department of Breast Surgery, Chris O'Brien Lifehouse, Camperdown, New South Wales, Australia. [17]Royal Prince Alfred Institute of Academic Surgery, University of Sydney, Sydney, New South Wales, Australia. [18]Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [19]Shanghai Institute of Immunology, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [20]Translational Immunology Institute, SingHealth Duke-NUS Academic Medical Centre, Singapore, Singapore. [21]New South Wales Health Pathology, Department of Anatomical Pathology, St George Hospital, Kogarah, New South Wales, Australia. [22]School of Medical Sciences, University of New South Wales, Sydney, New South Wales, Australia. [23]Faculty of Medicine & Health Sciences, Western Sydney University, Campbelltown, New South Wales, Australia. [24]University of New South Wales Cellular Genomics Futures Institute, University of New South Wales, Sydney, New South Wales, Australia. [25]Sydney Medical School, Sydney University, Sydney, New South Wales, Australia. [26]These authors contributed equally: Sunny Z. Wu, Ghamdan Al-Eryani, Daniel Lee Roden. ✉e-mail: a.swarbrick@garvan.org.au

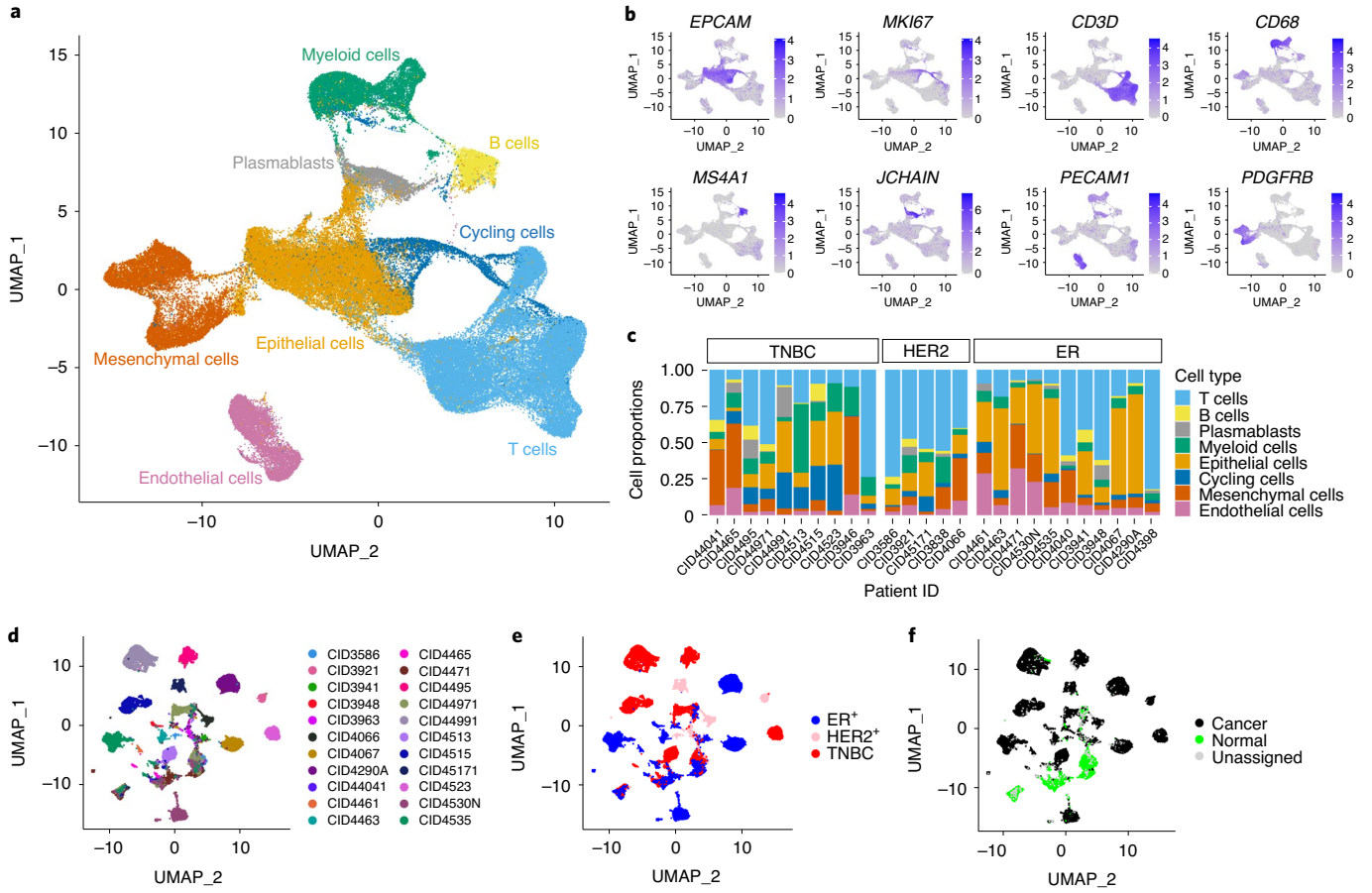


Fig. 1 | Cellular composition of primary breast cancers and identification of malignant epithelial cells. a, UMAP visualization of 130,246 cells analyzed by scRNA-seq and integrated across 26 primary breast tumors. Clusters were annotated for their cell types as predicted using canonical markers and signature-based annotation using Garnett. **b**, Log-normalized expression of markers for epithelial cells (*EPCAM*), proliferating cells (*MKI67*), T cells (*CD3D*), myeloid cells (CD68), B cells (*MS4A1*), plasmablasts (*JCHAIN*), endothelial cells (*PECAM1*) and mesenchymal cells (fibroblasts/perivascular-like cells; *PDGFRB*). **c**, Relative proportions of cell types highlighting a strong representation of the major lineages across tumors and clinical subtypes. **d**–**f**, UMAP visualization of all epithelial cells, from tumors with at least 200 epithelial cells, colored by tumor (**d**), clinical subtype (**e**) and inferCNV classification (**f**).

Breast cancers are diverse cellular microenvironments, whereby heterotypic interactions are important in defining disease etiology and response to treatment[7,8]. While breast cancers are generally considered to have a low mutational burden and immunogenicity, there is evidence that immune activation is pivotal in a subset of patients. For instance, the presence of tumor-infiltrating lymphocytes (TILs) is a biomarker for good clinical outcome and complete pathological response to neoadjuvant chemotherapy[9]. In contrast, tumor-associated macrophages (TAMs) are often associated with poor prognosis[10] and are recognized as important emerging targets for cancer immunotherapy[11–13].

Mesenchymal cells have also emerged as important regulators of the malignant phenotype, chemotherapy response[7] and antitumor immunity[14,15]. However, progress has been impeded by lack of a clear cellular taxonomy (recently reviewed in Sahai et al.[16]). Recent studies of cancer-associated fibroblasts (CAFs) identified two polarized states defined by extracellular matrix (ECM) production or inflammatory secretomes[17–19]. The relationship of these distinct cellular subsets with each other, with other cells in the tumor microenvironment (TME), and with disease status and progression is still to be elucidated in breast tumors.

Our understanding of the cellular heterogeneity and tissue architecture of human breast cancers has been largely derived from histology, bulk sequencing, low dimensionality hypothesis-based studies and experimental model systems. Single-cell RNA sequencing (scRNA-seq) offers remarkable new opportunities to systematically describe the cellular landscape of tumors[20,21] and reveal new insights into cell biology, disease etiology and drug response. Several studies have successfully applied scRNA-seq to selected populations in human breast tumors to reveal a continuum of differentiation states within TILs[22], a role for tissue-resident CD8 cells in TNBC[23] and chemoresistance of neoplastic cells in TNBC[24]. Recent studies have used mass cytometry with panels of antibody markers to analyze millions of cells from hundreds of patients to interrogate breast cancer cell types and ecosystems[25,26]. Therefore, a more detailed transcriptional atlas of breast tumors at high molecular resolution, representative of all subtypes and cell types, is required to further define the taxonomy of the disease, identify heterotypic cellular interactions and determine cellular differentiation events. Just as importantly, data systematically mapping the spatial transcriptomic architecture of breast tumors, which can determine how cells in the TME are organized as functional units, are scarce.

Results

A high-resolution cellular landscape of human breast cancers. To elucidate the cellular architecture of breast cancers, we analyzed 26 primary tumors, including 11 ER$^+$, 5 HER2$^+$ and 10 TNBCs, by scRNA-seq (Supplementary Table 1). In total, 130,246 single cells

passed quality control (Extended Data Fig. 1a–d) and were annotated using canonical lineage markers (Fig. 1a,b). These high-level annotations were further confirmed using published gene signatures[27–29]. All major cell types were represented across all tumors and clinical subtypes (Fig. 1c). As previously reported in other cancers[30,31], uniform manifold approximation and projection (UMAP) visualization showed a clear separation of epithelial cells by tumor, although three clusters contained cells from multiple patients and subtypes (Fig. 1d,e), which were identified as normal breast epithelial cells (Fig. 1f). In contrast, UMAP visualization of stromal and immune cells across tumors clustered together without batch correction (Extended Data Fig. 1e,f). Since breast cancer is largely driven by DNA copy number changes[32], we estimated single-cell copy number variant (CNV) profiles using inferCNV[31] to distinguish neoplastic from normal epithelial cells (Fig. 1f). Within neoplastic populations, substantial levels of large-scale genomic rearrangements were observed (Extended Data Fig. 1g and Supplementary Table 2). This revealed patient-unique copy number changes and those commonly seen in breast cancers, such as chr1q gain in luminal cancers and chr5q loss in basal-like breast cancers[32].

SCSubtype: intrinsic subtyping for scRNA-seq data. Since unsupervised clustering could not be used to find recurring neoplastic cell gene expression features between tumors, we asked whether we could classify cells using the established PAM50 method. Due to the inherent sparsity of single-cell data, we developed a scRNA-seq-compatible method for intrinsic molecular subtyping. We constructed 'pseudobulk' profiles from scRNA-seq for each tumor and applied the PAM50 centroid predictor. To identify a robust training set, we used hierarchical clustering of the pseudobulk samples with The Cancer Genome Atlas (TCGA) dataset of 1,100 breast tumors using an approximate 2,000-gene intrinsic breast cancer gene list[3] (Extended Data Fig. 2a,b). Training samples were selected from those with concordance between pseudobulk PAM50 subtype calls and TCGA clusters (Supplementary Table 3).

For each PAM50 subtype within the training dataset, we performed pairwise integrations of tumor cells and differential gene expression to identify 4 sets of genes that would define our single-cell-derived molecular subtypes (89 genes, Basal_SC; 102 genes, HER2E_SC; 46 genes, LumA_SC; 65 genes, LumB_SC). We defined these genes as the 'SCSubtype' gene signatures (Fig. 2a, Extended Data Fig. 2c and Supplementary Table 4). Only four of these genes showed overlap with the original PAM50 gene list (*ACTR3B*, *KRT14*, *ERBB2*, *GRB7*). A subtype call for a given cell was based on the maximum SCSubtype score. An overall tumor subtype was then assigned based on the majority cell subtype. This approach showed 100% agreement with the PAM50 pseudobulk calls in the 10 training set samples and 66% agreement in the test set samples (Extended Data Fig. 2d and Supplementary Table 3). Of the three test set disagreements, two were LumA versus LumB, which are related profiles that may have been hard to distinguish with a limited sample size, and the third was a metaplastic TNBC sample, which is a histological subtype not included in the original PAM50 training or test datasets.

As another means of assessing the accuracy of SCSubtype, we performed 'true bulk' whole-transcriptome RNA-seq on 16 matching tumors in our scRNA-seq cohort. We observed concordance between the majority SCSubtype calls and bulk tumor RNA-seq profiles in 12 of 16 tumors, including 7 of the 8 matching training set tumors (Supplementary Table 3). We also clustered the bulk RNA-seq data with TCGA, confirming that 14 of the samples clustered with their pseudobulk profiles (Extended Data Fig. 2a–b). These results highlight the strong concordance between our three subtyping methods when applied across bulk and scRNA-seq datasets.

SCSubtype revealed that 13 out of 20 samples had less than 90% of neoplastic cells falling under 1 molecular subtype, while only 1 tumor (CID3921; HER2E) showed a completely homogenous molecular subtype (Fig. 2b). In some luminal and HER2E tumors, SCSubtype predicted small numbers of basal-like cells, which was validated by immunohistochemistry (IHC) in two ER$^+$ cases that showed small pockets of morphologically malignant cells that were negative for ER and positive for cytokeratin-5 (CK5), a basal cell marker, among otherwise ER$^+$ tumor cells (Fig. 2c). The utility of SCSubtype was further demonstrated by its ability to correctly assign a low cellularity lobular carcinoma (10% neoplastic cells; CID4471), evident both by histology and inferCNV (Supplementary Table 2), as a mixture of mostly LumB and LumA cells (Fig. 2b and Extended Data Fig. 2d), which is consistent with the clinical IHC result. Bulk and pseudobulk RNA-seq incorrectly assigned CID4471 as normal-like (Supplementary Table 3).

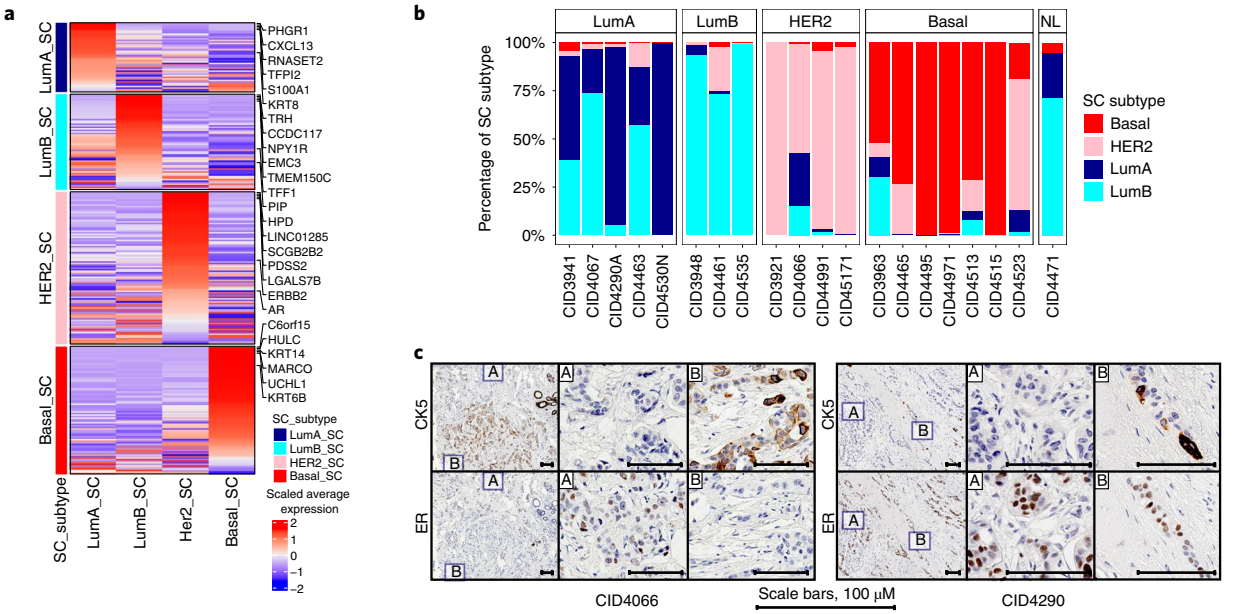
To further validate SCSubtype, we calculated the degree of epithelial cell differentiation (DScore)[33] and proliferation[34], both of which are independently associated with the molecular subtype of each cell. Basal_SC cells tended to have low DScores and high proliferation scores whereas LumA_SC cells showed high DScores and low proliferation scores (Fig. 2d and Extended Data Fig. 2e), as observed across PAM50 subtypes in TCGA (Extended Data Fig. 2f).

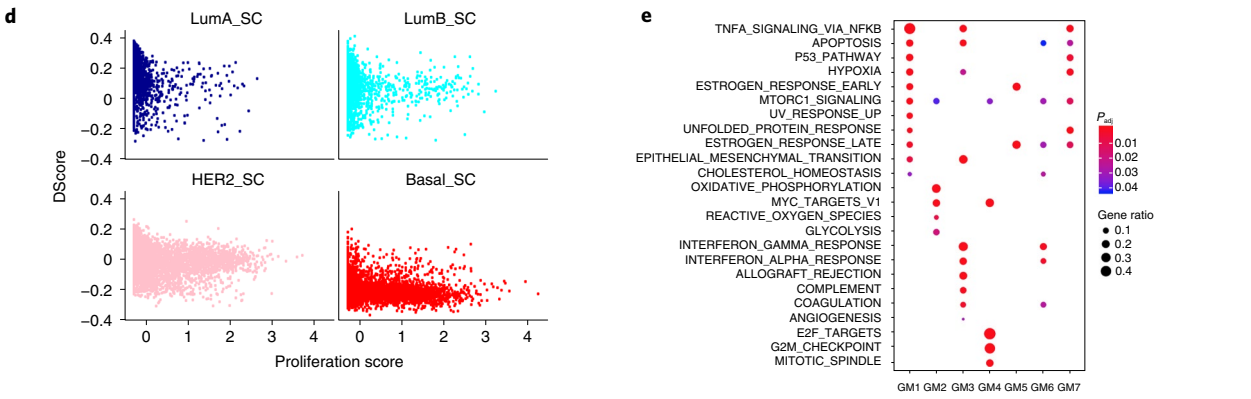
Recurrent gene modules driving neoplastic cell heterogeneity. The previous method relied on a priori knowledge of a 'bulk' molecular subtype to develop a classifier. To complement this, we investigated the biological pathways driving intratumor transcriptional heterogeneity (ITTH) in an unsupervised manner, using integrative clustering of tumors with at least 50 neoplastic cells, to generate 574 gene signatures of ITTH. These gene signatures identified seven robust groups, 'gene modules' (GMs), based on their Jaccard similarity (Extended Data Fig. 3a). Each GM was defined with 200 genes that had the highest frequency of occurrence across the ITTH gene signatures and individual tumors (Supplementary Table 5), minimizing the contribution of a single tumor to any particular module.

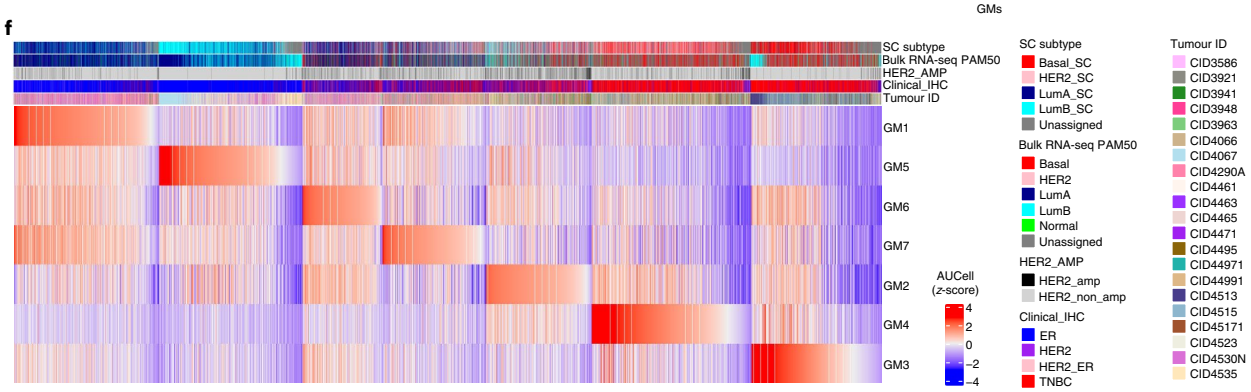
Gene-set enrichment identified shared and distinct functional features of these GMs (Fig. 2e). GM4 was uniquely enriched for hallmarks of cell cycle and proliferation (for example, E2F_TARGETS),

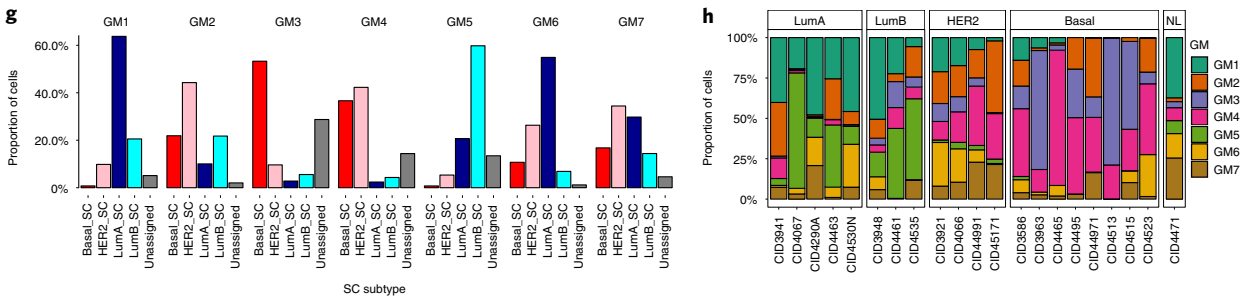
Fig. 2 | Identifying drivers of neoplastic breast cancer cell heterogeneity. a, Heatmap showing the average expression (scaled) of all cells for the SCSubtype gene signatures assigned to each of the four SCSubtypes. The top five most highly expressed genes in each subtype are shown; selected others are highlighted. **b**, Percentage of neoplastic cells in each tumor that are classified as each of the SCSubtypes. Tumor samples are grouped according to their Allcells-Pseudobulk classifications. NL, normal-like. **c**, Representative images of CK5 (top) and ER (bottom) IHC from two tumors (CID4066, left; CID4290, right) with intrinsic subtype heterogeneity from **b** ($n = 24$ breast tumors analyzed). Left: whole-tissue sections with two regions of interest labeled (A and B). Middle: CK5$^-$/ER$^+$ areas (insert A). Right: CK5$^+$/ER$^-$ areas (insert B). Scale bar, 100 µm. **d**, Scatter plots of the proliferation and differentiation scores (DScores) of each neoplastic cell. Individual cancer cells are colored and grouped based on the SCSubtype calls. All pairwise comparisons between cells from each SCSubtype were significantly different (Wilcoxon signed-rank test $P < 0.001$) for both proliferation and DScores. **e**, Gene-set enrichment using ClusterProfiler of the 200 genes in each of the GMs (GM1–GM7). Significantly enriched (Benjamini–Hochberg-adjusted $P < 0.05$) gene sets from the MSigDB HALLMARK collection are shown. **f**, Scaled signature scores of each of the seven intratumor transcriptional heterogeneity GMs (rows) across all individual neoplastic cells (columns). Cells are ordered based on the strength of the GM signature score. **g**, Proportion of cells assigned to each of the SCSubtypes grouped according to GM. **h**, Percentage of neoplastic cells assigned to each of the seven GMs.

driven by genes including *MKI67*, *PCNA* and *CDK1*. GM3 was predominantly enriched for hallmarks of interferon response (*IFITM1/2/3*, *IRF1*), antigen presentation (*B2M*; *HLA-A/B*) and epithelial–mesenchymal transition (EMT; *VIM*, *ACTA2*). GM1 and GM5 showed characteristics of estrogen response pathways, while GM1 was also enriched for hypoxia, tumor necrosis factor-α and









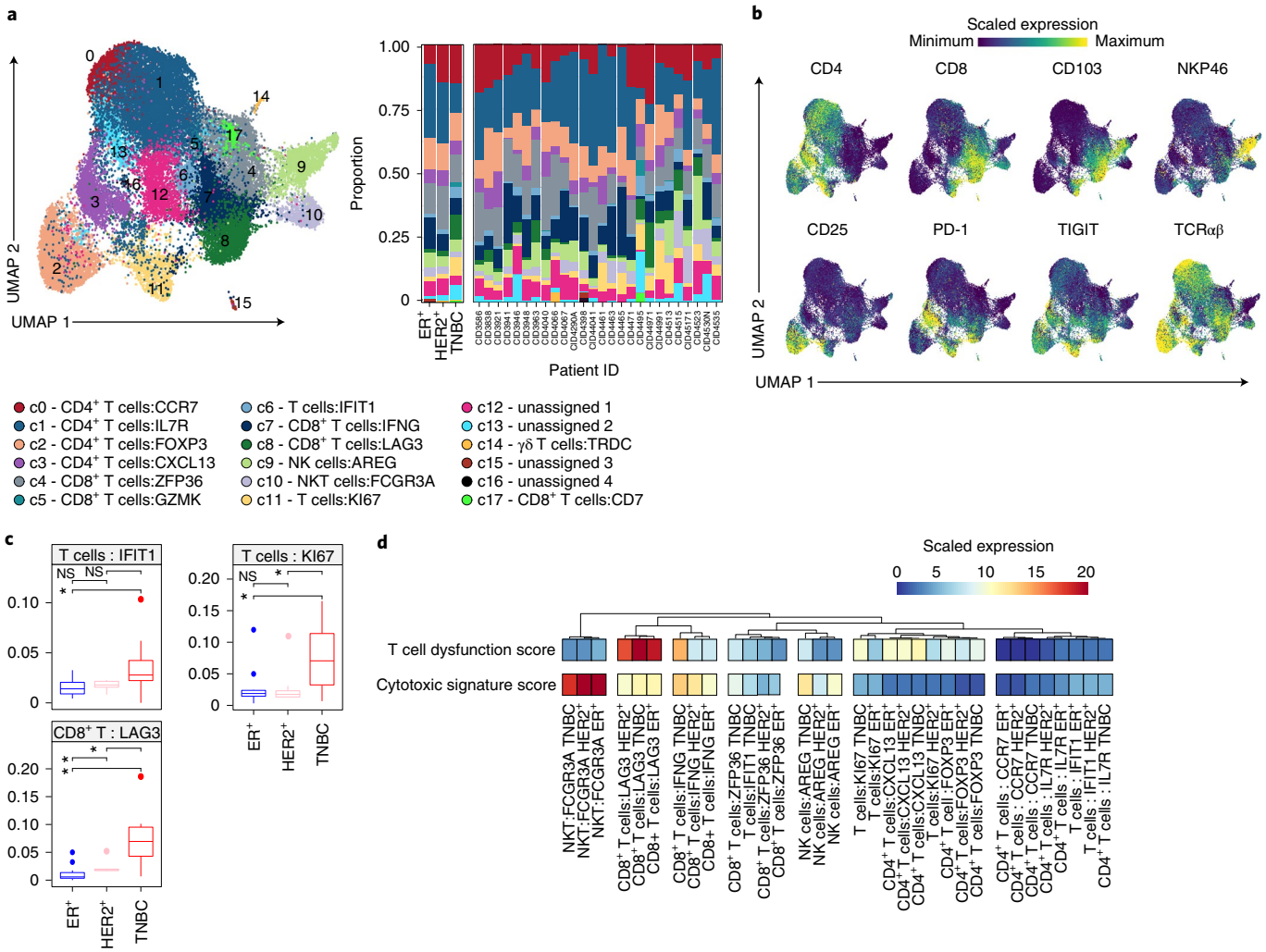


Fig. 3 | T cell and innate lymphoid cell landscape of breast cancers. a, Reclustering T cells and innate lymphoid cells and their relative proportions across tumors and clinical subtypes ($n = 35{,}233$ cells from 26 tumors). **b**, Imputed CITE-seq protein expression values for selected markers and checkpoint molecules. **c**, Pairwise *t*-test comparisons revealing the significant enrichment of T cells:IFIT1, T cells:Ki67 and CD8$^+$ T cells:LAG3 in TNBC tumors ($n = 26$; 11 TNBCs, 10 ER$^+$ and 5 HER2$^+$). The box plots depict the first and third quartiles as the lower and upper bounds, respectively. The whiskers represent 1.5× the IQR and the center depicts the median. Statistical significance was determined using a two-sided *t*-test in a pairwise comparison of means between groups, with *P* values adjusted using the Benjamini–Hochberg procedure. $^*P < 0.05$, $P < 0.01$, $^*P < 0.001$ and $^{****}P < 0.0001$. NS, not significant. **d**, Cluster-averaged dysfunctional and cytotoxic effector gene signature scores in T cells and innate lymphoid cells stratified by clinical subtypes.

p53 signaling and apoptosis. Similar functional associations were also seen when correlating signature scores across all neoplastic cells (Extended Data Fig. 3b).

For each neoplastic cell, we calculated signature scores for the seven GMs and used hierarchical clustering to identify cellular correlations (Extended Data Fig. 3c). This clearly separated neoplastic cells into groups, reducing the large intertumor variability seen in Fig. 1d–f. We assigned each neoplastic cell to a module using the maximum of the scaled scores (Extended Data Fig. 3d). Some modules were associated with SCSubtype calls, whereas others displayed more diverse subtype associations (Fig. 2f,g and Extended Data Fig. 3e,f). Cells assigned to GM1 and GM5 were predominantly enriched for the luminal subtype, whereas GM1 was almost exclusively composed of LumA cells and GM5 was mostly composed of LumB cells. Since proliferative cells were classified separately, as GM4, this suggests that there were subsets of cells within LumA tumors with unique properties not found in LumB tumors. Finally, we used the GM-based cell state assignments to get a view into the intratumor heterogeneity of the neoplastic cells. Similar to SCSubtype (Fig. 2b), we saw evidence for cellular heterogeneity that broadly aligned with, but was not constrained by, the tumor subtype (Fig. 2h). SCSubtype and GM analysis provide complementary new approaches to classifying neoplastic ITTH and provide further evidence that cancer cells manifest diverse phenotypes within most tumors.

The immune milieu of breast cancer. To examine the immune milieu of breast tumors at high resolution, we reclustered immune cells to identify T cells and innate lymphoid cells, myeloid cells, B cells and plasmablasts (Supplementary Table 6). We performed immunophenotyping using cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq)[35] to four samples and performed anchor-based integration to transfer protein expression levels to the remaining cases[36], which revealed a high correlation to experimentally measured values (Extended Data Fig. 4).

Lymphocytes and innate lymphoid cells. We identified 18 T cell and innate lymphoid clusters across patients (Fig. 3a). CD4 clusters consisted of *FOXP3*$^+$ regulatory T (T_{reg}) cells marked by CD25

protein expression (CD4+ T cells:*FOXP3*/c2), follicular helper T (T_{FH}) cells (*CXCL13*, *IL21* and *PDCD1*; CD4+ T cells:*CXCL13*/c3), naive/central memory CD4+ (CD4+ T cells:*CCR7*/c0) and a type 1 helper T (T_H1) CD4 effector memory T (T_{EM}) cluster (CD4+ T cells:*IL7R*/c1) (Fig. 3b and Extended Data Fig. 5a). Of the five CD8 clusters, three consisted of a cluster with high expression of inhibitory checkpoint molecules including *LAG3*, *PDCD1* and *TIGIT* (CD8+ T cells:*LAG3*/c8), *PDCD1*-low CD8+ T cells that expressed relatively high levels of *IFNG* and *TNF* (CD8+ T cells:*IFNG*/c7) and chemokine-expressing T cells (CD8+ T cells:*ZFP36*/c4) (Extended Data Fig. 5a). Two additional clusters driven by a type 1 interferon (IFN-I) signature (*SG15*, *IFIT1* and *OAS1*; T cells:*IFIT1*/c6) and proliferation (T cells:*MKI67*/c11) were identified, both consisting of CD4+ and CD8+ T cells. We also identified natural killer (NK) cells (NK cells:*AREG*/c9) and natural killer T (NKT)-like cells (NKT cells:*FCGR3A*/c10) by their expression of αβ T cell receptor and NK markers (*KLRC1*, *KLRB1*, *NKG7)* (Fig. 3b and Extended Data Fig. 5a).

Consistent with the enrichment of TILs and CD8+ T cells in TNBC[37], the T cell clusters *IFIT1*/c6, *LAG3*/c8 and *MKI67*/c11 made up a higher proportion in the TNBC samples (Fig. 3c). These clusters had qualitative differences between clinical subtypes, with CD8+ T cells from both the *LAG3*/c8 and *IFNG*/c7 clusters possessing substantially higher dysfunction scores[38] in TNBC cases (Fig. 3d and Extended Data Fig. 5b,c). Furthermore, luminal and HER2+ tumors tended to have checkpoint molecule expression distinct from TNBC (Fig. 4f and Extended Data Fig. 5d). Notably, the *LAG3*/c8-exhausted CD8 subset in TNBCs had significantly higher expression of PD-1 (*PDCD1*), LAG3 and the ligand–receptor pair of CD27 and CD70, known to enhance T cell cytotoxicity[39] (Fig. 4f and Extended Data Fig. 5e). We examined the expression of *PDCD1*, *CD27*, *LAG3* and *CD70* in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)[40] cohort, which were consistently enriched in basal-like and HER2+ subtypes (Extended Data Fig. 5f). When we examined a wider list of immune checkpoint molecules across the entire dataset using unsupervised hierarchical clustering (Extended Data Fig. 6), differences in checkpoint molecule expression among clinical subtypes were more apparent, including on non-immune cells such as CAFs. These data provide insights into the immunotherapeutic strategies most appropriate for each subtype of disease.

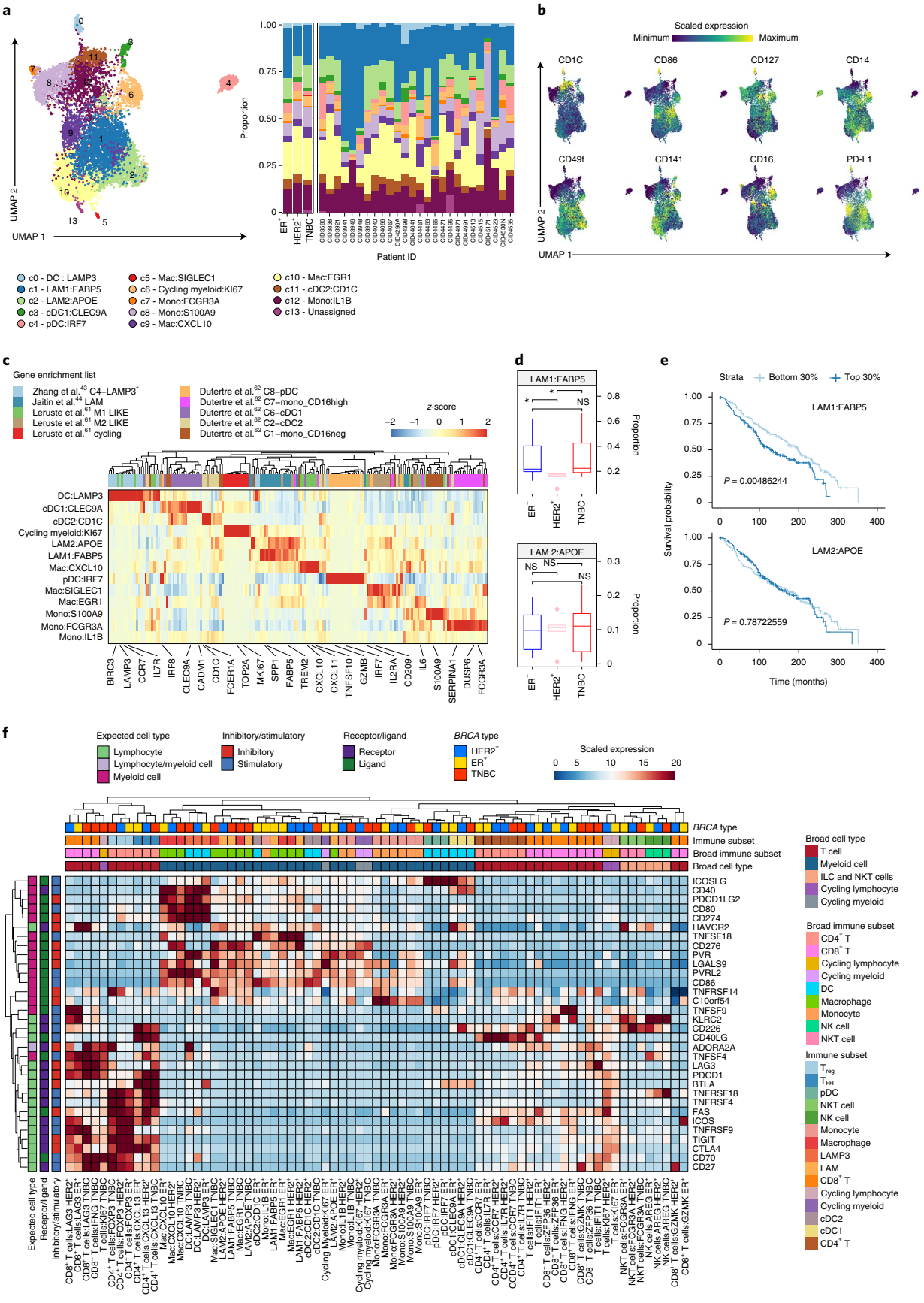
When we reclustered B cells, we observed two major subclusters (naive and memory), with plasmablasts forming a separate cluster (Extended Data Fig. 7a,b). The additional subclusters seemed largely driven by B cell antigen receptor-specific gene segments rather than variable biological gene expression programs.

Myeloid cells. Myeloid cells formed 13 clusters that could be identified in all tumors at varying frequencies (Fig. 4a). No granulocytes were detected, probably due to their sensitivity to tumor dissociation protocols and their low abundance[22,41,42]. Monocytes formed three clusters: Mono:*IL1B*/c12; Mono:*S100A9*/c8; and Mono:*FCGR3A*/c7. The Mono:*FCGR3A* population formed a small distinct cluster characterized by high CD16 protein expression (Fig. 4b,c). We identified conventional dendritic cells (cDCs) that expressed either *CLEC9A* (cDC1:*CLEC9A*/c3) or *CD1C* (cDC2:*CD1C*/c11), plasmacytoid DCs (pDCs) that expressed *IRF7* (pDC:*IRF7*/c4) and a *LAMP3* high DC population[43] (DC:*LAMP3*/c0), which was previously not reported in single-cell studies of breast cancer (Fig. 4c and Extended Data Fig. 7d). Macrophages formed six clusters, including a cluster (Mac:*CXCL10*/c9) with features previously associated with an 'M1-like' phenotype and two clusters (Mac:*EGR1*/c10 and Mac:*SIGLEC1*/c5) resembling the 'M2-like' phenotype, all of which bear some resemblance to TAMs previously described in breast cancers (Extended Data Fig. 7c)[10]. Notably, we identified two new macrophage populations (LAM1:*FABP5*/c1 and LAM2:*APOE*/c2) outside of the conventional M1/M2 classification that comprised 30–40% of total myeloid cells (Fig. 4a–c). These cells bear close transcriptomic similarity to a recently described population of lipid-associated macrophages (LAMs) that expand in obese mice and humans[44], including high expression of *TREM2* and lipid/fatty acid metabolic genes such as *FABP5* and *APOE* (Fig. 4c and Extended Data Fig. 7d,e). LAM1/2 uniquely expressed *CCL18*, which encodes a chemokine with roles in immune regulation and tumor promotion[45]. We observed a substantially reduced proportion of LAM1:*FABP5* cells in HER2+ tumors (Fig. 4d and Extended Data Fig. 7f), suggesting that unique features of tumor genomics or microenvironment regulate LAM1/2 fate. Survival analysis using the METABRIC[40] cohort showed that the LAM1:*FABP5* signature correlates with worse survival (Fig. 4e). While the RNA encoding PD-L1 (*CD274*) and PD-L2 (*PDCD1LG2*) was highly coexpressed by the Mac:*CXCL10* and DC:*LAMP3* myeloid populations (Fig. 4f), analysis of the CITE-seq data demonstrated a broader distribution of PD-L1 and PD-L2 protein expression across the Mac:*CXCL10*, LAM1:*FABP5*, LAM2:*APOE* and DC:*LAMP3* clusters (Fig. 4b and Extended Data Fig. 7g), highlighting LAM1/2 as important sources of immunoregulatory molecules.

Stromal subclasses resemble diverse differentiation states. In the stromal compartment, we identified three major cell types (Fig. 5a,b and Extended Data Fig. 8a) including CAFs (*PDGFRA* and *COL1A1;* Fig. 5c,d), perivascular-like (PVL) cells (*MCAM*/CD146, *ACTA2* and *PDGFRB;* Fig. 5e,f), endothelial cells (*PECAM1*/CD31 and *CD34;* Fig. 5g,h), plus two smaller clusters of lymphatic endothelial cells (*LYVE1*) and cycling PVL cells (*MKI67*)[15]. Pseudotime trajectory analysis using Monocle 2[46] revealed five CAF states (Fig. 5c and Extended Data Fig. 8b,c). State 1 (referred to as s1 from this point onward) had features of mesenchymal stem cells (MSCs) and inflammatory-like CAFs (iCAFs), with high expression of stem cell markers (*ALDH1A1*, *KLF4* and *LEPR*) and pathways related to chemoattraction and complement cascades (*CXCL12* and *C3*) (Extended Data Fig. 8d,e). The expression of these markers decreased as cells transitioned toward differentiated states s4 and s5, which resembled myofibroblast-like CAF states through the increased expression of *ACTA2* (αSMA), *TAGLN*, *FAP* and *COL1A1* (ref. [15]) and the enrichment of ECM-related pathways. Previously

Fig. 4 | Myeloid landscape of breast cancers. a, Reclustered myeloid cells and their relative proportions across tumors and clinical subtypes ($n = 9,678$ cells from 26 tumors). **b**, Imputed CITE-seq expression values for canonical markers and checkpoint molecules across myeloid clusters. **c**, Cluster-averaged expression of various published gene signatures acquired from independent studies used for myeloid cluster annotation. Selected genes of interest from each signature are listed. References[43,44,61,62] are cited in this panel. **d**, Proportions of LAM1:*FABP5* and LAM2:*APOE* ($n = 26$; 11 TNBCs, 10 ER+ and 5 HER2+) across clinical subtypes. The box plots depict the first and third quartiles as the lower and upper bounds, respectively. The whiskers represent 1.5× the IQR and the center depicts the median. Statistical significance was determined using a two-sided *t*-test in a pairwise comparison of means between groups, with *P* values adjusted using the Benjamini–Hochberg procedure. $^*P < 0.05$. **e**, Kaplan–Meier plots showing the associations between LAM1:*FABP5* or LAM2:*APOE* with overall survival in the METABRIC cohort (top and bottom 30%, $n = 180$ per group). *P* values were calculated using the log-rank test. **f**, Cluster-averaged gene expression of clinically relevant immunotherapy targets. Clusters are grouped by breast cancer clinical subtype and immune cell type annotations. Genes are grouped as receptor (purple) or ligand (green), inhibitory (red) or stimulatory status (blue) and expected major lineage cell types known to express the gene (lymphocyte, green; myeloid, pink; both, light purple).

reported iCAF and myofibroblast-like CAF signatures from pancreatic ductal adenocarcinoma[19] were predominantly enriched in CAF s1 and s5, respectively (Extended Data Fig. 8f). No CAF states were enriched for antigen presentation CAF signatures; however, selected antigen presentation CAF markers *CD74*, *CLU* and *CAV1* were broadly expressed across all stromal cells (Extended Data Fig. 8g).



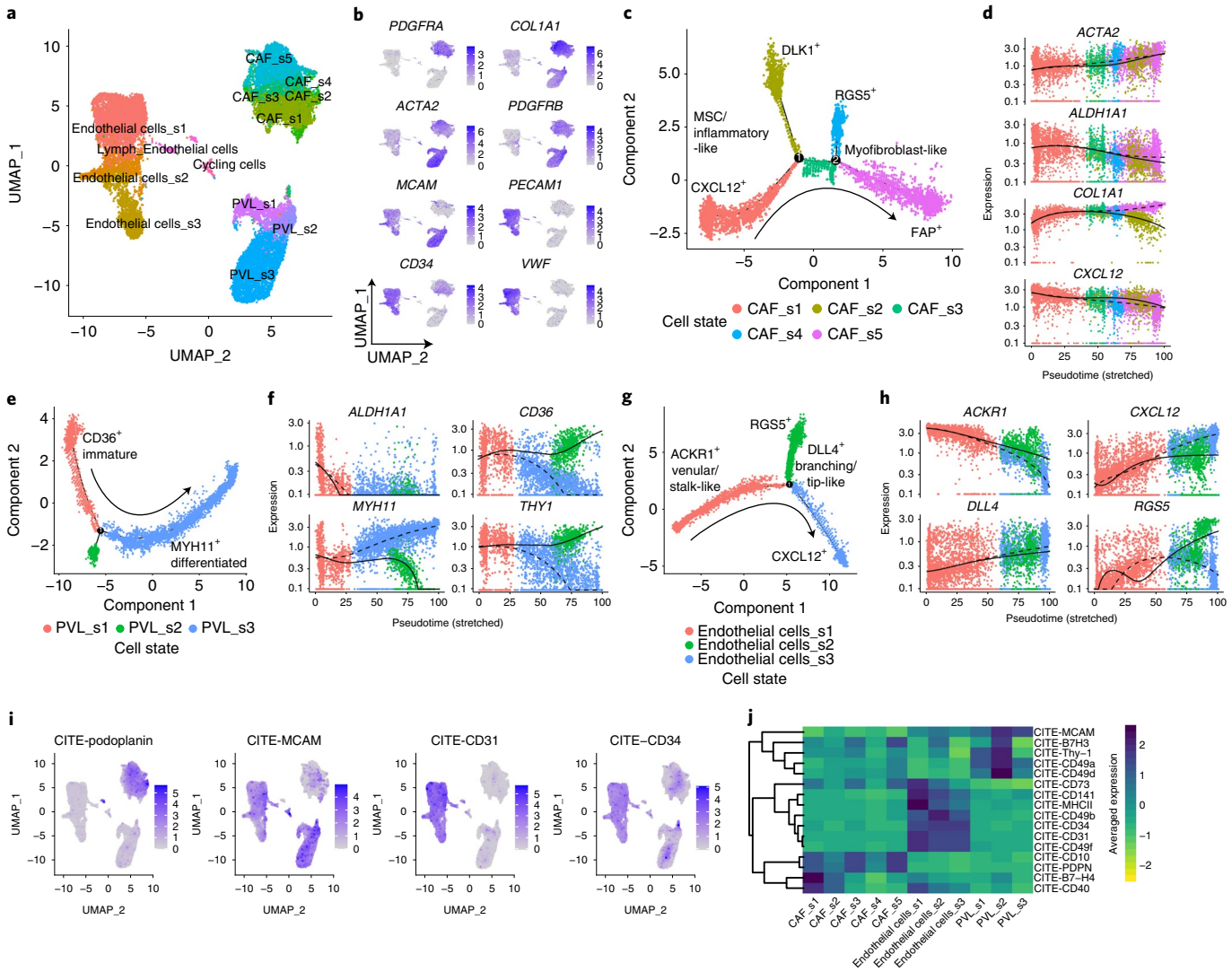


Fig. 5 | Transcriptional profiling and phenotyping of diverse mesenchymal differentiation states across breast cancers. a, UMAP visualization of reclustered mesenchymal cells, including CAFs (6,573 cells), PVL cells (5,423 cells), endothelial cells (7,899 cells), lymphatic endothelial cells (203 cells) and cycling PVL cells (50 cells). Cell substates were defined using pseudotemporal ordering with Monocle 2 (as in **c**–**h**). **b**, Feature plots of canonical markers for CAFs (*PDGFRA*, *COL1A1*, *ACTA2*, *PDGFRB*), PVLs (*ACTA2*, *PDGFRB* and *MCAM*) and endothelial cells (*PECAM1*, *CD34* and *VWF*). The UMAP axes correspond to **a**. **c**–**h**, Cell states and the expression of genes that change as a function of pseudotime for CAFs (**c**–**h**), PVL cells (**e**,**f**) and endothelial cells (**g**,**h**). **c**,**d**, Five states of CAFs: CAF_s1 and CAF_s2 both resemble MSCs (*ALDH1A1*) and iCAF states (*CXCL12*); CAF_s2 was distinct from CAF_s1 by *DLK1*; CAF_s4 and CAF_s5 resemble myofibroblast-like states (myofibroblast-like CAFs; *ACTA2*), which were enriched for ECM genes (*COL1A1*); transitioning CAF_s3 shared features of both MSC/iCAFs and myofibroblast-like CAFs. **e**,**f**, Three PVL states: PVL_s1 and PVL_s2 resemble progenitor and immature states (immature PVLs; *ALDH1A1*); PVL_s3 resembles a contractile and differentiated state (*MYH11*). **g**,**h**, Three endothelial states: s1, which resembles a venular stalk-like state (*ACKR1*) and two tip-like states (*DLL4*) s2 and s3, which are distinguished by *RGS5* and *CXCL12*, respectively. **i**, Feature plots of imputed CITE-seq antibody-derived tag protein levels for canonical markers of CAFs (podoplanin), PVL cells (CD146/*MCAM*) and endothelial cells (CD31 and CD34). The UMAP coordinates correspond to those in **a**. **j**, Heatmap of cluster-averaged imputed CITE-seq values for additional cell-surface markers and functional molecules.

For PVL cells, we identified three states (Fig. 5e). PVL s1 and s2 expressed markers related to stem cells, immature pericytes (*PDGFRB*, *ALDH1A1*, *CD44*, *CSPG4*, *RGS5* and *CD36*) and adhesion molecules (*ICAM1*, *VCAM1* and *ITGB1*) (Extended Data Fig. 8d)[47]. They were further enriched for pathways related to receptor binding and platelet-derived growth factor activity (Extended Data Fig. 8e). The branching of s2 was defined by *RGS5*, *CD248* and *THY1*. Consistent with gene expression, CITE-seq revealed an enrichment of cell-surface CD90 (*THY1*) and the integrin molecules CD49a and CD49d in early PVL states s1 and s2 (Fig. 5i,j). The expression of these markers decreased as cells transitioned to PVL s3, which was enriched for contractile related genes (*MYH11* and *ACTA2*) (Fig. 5f) and pathways related to a smooth muscle phenotype. PVL states were modestly enriched for myofibroblast-like CAF gene signatures (Extended Data Fig. 8f); their shared expression of the CAF marker *ACTA2* suggest that PVL s3 cells have historically been misclassified in IHC assays as CAFs.

We identified three endothelial states (Fig. 5g). Endothelial s1 resembled stalk-like and venular endothelial cells (*ACKR1*, *SELE* and *SELP*)[48], enriched for pathways and genes related to cell adhesion (*ICAM1* and *VCAM1*) and antigen presentation/major histocompatibility complex (MHC) (*HLA-DRA*) (Extended Data

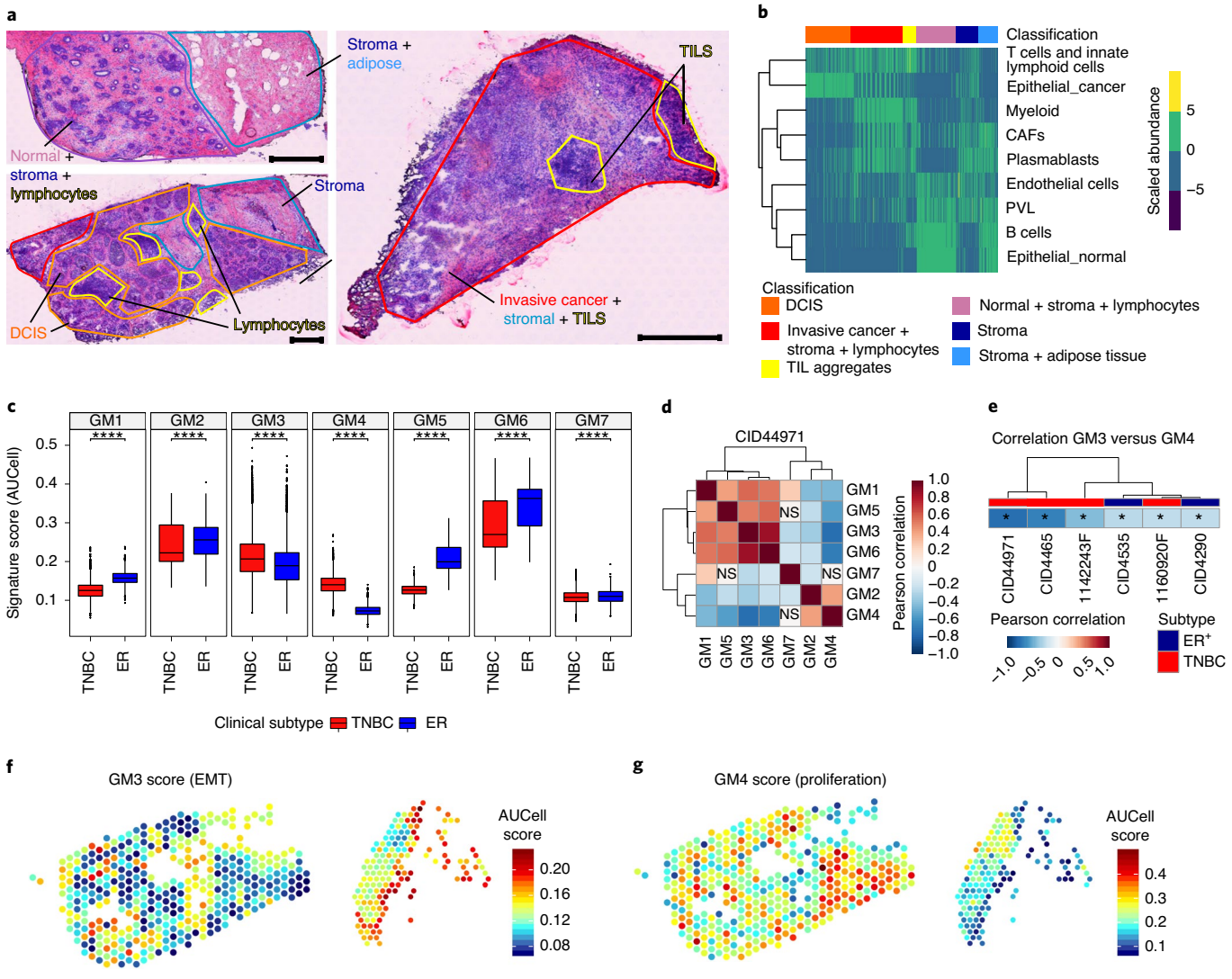


Fig. 6 | Mapping breast cancer heterogeneity using spatially resolved transcriptomics. a, Complete H&E images of all three tissue regions analyzed using Visium for the sample TNBC CID44971. Pathological annotation of morphological regions into distinct categories including normal ductal (purple), stromal and adipose (blue), lymphocyte aggregates (yellow), ductal carcinoma in situ (DCIS) (orange) and invasive cancer (red). Black scale bars, 500 µm. **b**, Deconvolution of the major cell type lineages in TNBC CID44971. Values signify the scaled cell type abundances per spots (columns) and are grouped by pathology annotation as in **a**. **c**, Box plot of GM scores grouped by clinical subtype across the six cases ($n = 11{,}535$ spots from 4 TNBC tumors and 2 ER$^+$ tumors). Only cancer-filtered spots were used for this analysis. Signature scores were computed using the AUCell method. Statistical significance was determined using a two-sided *t*-test in a comparison of means between groups, with *P* values adjusted using the Benjamini–Hochberg procedure. The box plots depict the first and third quartiles as the lower and upper bounds, respectively. The whiskers represent 1.5× the IQR and the center depicts the median. $^{****}P < 0.0001$. **d**, Pearson correlation heatmap of breast cancer GMs in TNBC CID44971 (two-sided correlation coefficient, Benjamini–Hochberg-adjusted $P < 0.05$). Spots with high cancer epithelial abundances (>10%) were scored with GM signatures using AUCell. **e**, Negative correlation between GM3 (EMT) and GM4 (proliferation/cell cycle) across all cancer epithelial spots from six breast cancers analyzed by spatially resolved transcriptomics (two-sided correlation coefficient, $^*P < 0.05$). **f**,**g**, AUCell signature scores of GM3 (**f**) and GM4 (**g**) overlaid onto cancer epithelial spots in TNBC CID44971, as defined in the bottom left and right tissue sections in **a**.

Fig. 8d,e). These markers decreased along pseudotime as cells branched into two states, both with elevated expression of *DLL4*, a marker reported for endothelial tip-like cells (Fig. 5h)[49,50]. Endothelial s2 was distinguished by *RGS5* and *ESM1*, while s3 expressed regulators of cell migration and angiogenesis (*CXCL12* and *VEGFC*)[51]. Since angiogenesis is known to be a dynamic process involving the transition between endothelial stalk and tip cells[52,53], it is likely that these three states, defined by the markers *ACKR1*, *RGS5* and *CXCL12*, are dynamic and interconvertible. Similar CAF, PVL and endothelial cell states were identified across clinical subtypes and in three normal breast tissue samples (Extended Data Fig. 8h,i), suggesting they are probably resident cell types that undergo remodeling in the TME.

Spatially mapping breast cancer heterogeneity. To gain insights into the spatial organization of cell types, we performed spatially resolved transcriptomics on six samples ('local cohort') comprising two ER$^+$ (CID4535 and CID4290) and two TNBCs (CID44971 and CID4465) from our scRNA-seq cohort, and two additional TNBCs (1142243F and 1160920F) processed in an independent laboratory (Fig. 6a and Extended Data Fig. 9a). To deconvolute the cellular composition of each approximate 55-µM diameter spot, we

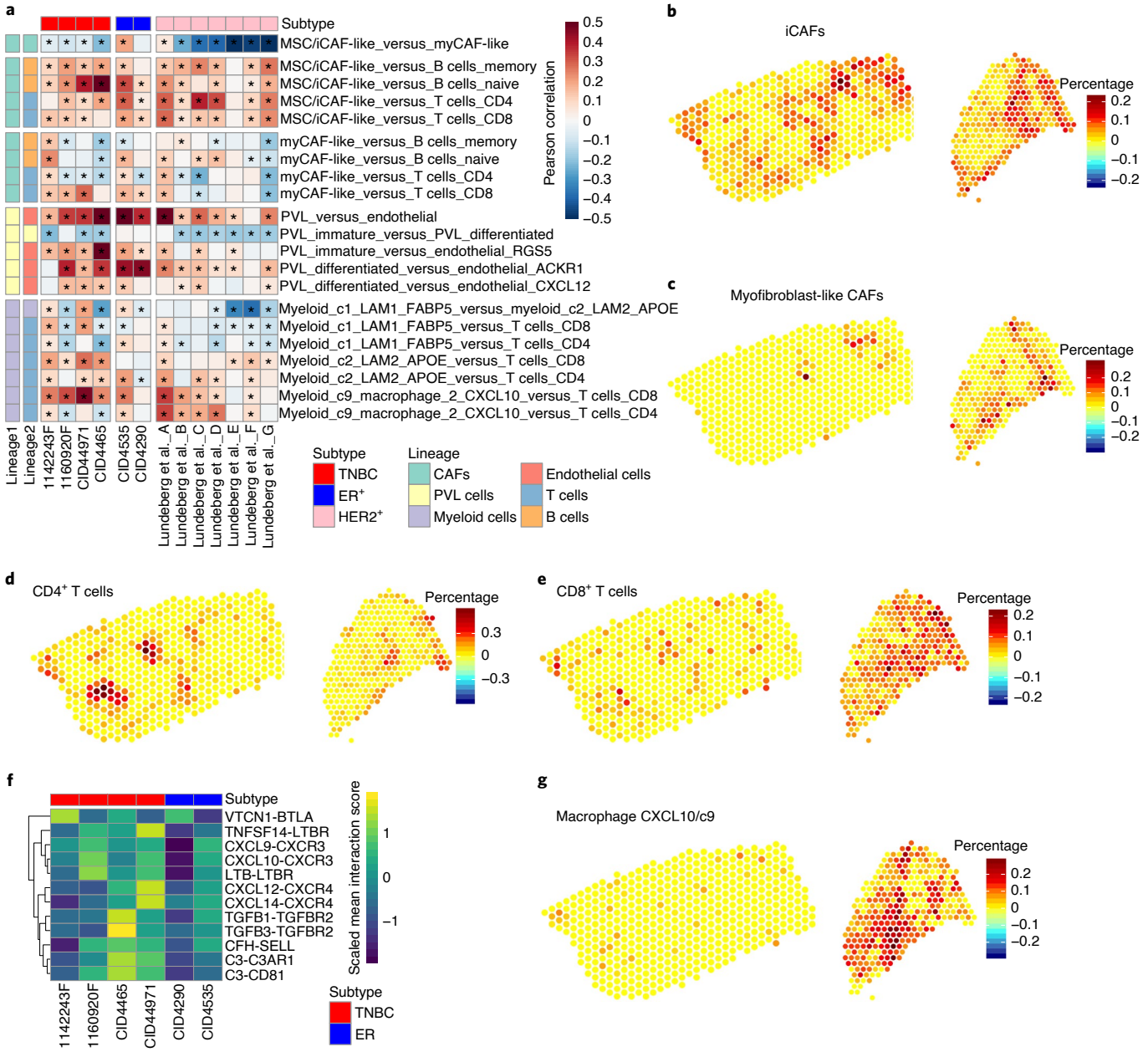


Fig. 7 | Spatially mapping new heterotypic cellular interactions. a, Heatmap of Pearson correlation values between subclasses of CAFs, PVL cells, endothelial cells, macrophage subsets and lymphocytes in 13 cases (two-sided correlation coefficient, *Benjamini–Hochberg-adjusted $P < 0.05$). Each tumor is stratified by the clinical subtype, including four TNBCs (blue), two ER$^+$ (red) analyzed in this study and seven HER2$^+$ (pink) cases from the Lundeberg et al. study[56]. **b**–**e**, Scaled deconvolution values for iCAFs (**b**), myofibroblast-like CAFs (**c**), CD4$^+$ (**d**) and CD8$^+$ T cells (**e**) overlaid onto tissue spots, as defined in the bottom left and right tissue sections in Fig. 6a. Representative TNBC case CID44971 is shown. **f**, Spatial proximity of selected CAF T cell signaling molecules. Heatmap of interaction scores for selected ligand–receptor pairs in the top 10% of tissue spots enriched for iCAFs and CD4$^+$/CD8$^+$ T cells. Only differentially expressed CAF ligands and T cell receptors detected by scRNA-seq using MAST were included. **g**, Scaled deconvolution values for macrophage *CXCL10*/c9 cells overlaid onto tissue spots as defined in Fig. 6a. Representative TNBC case CID44971 is shown.

applied a probabilistic model called Stereoscope[54] using clinical subtype-matched scRNA-seq data. Cell types were associated with their appropriate pathological annotation (Fig. 6b).

We earlier showed that GMs were enriched for distinct microenvironment-associated pathways and factors; thus, we hypothesized that GMs would be spatially organized in breast tumors. We selected locations in all six cases where cancer cells were identified by Stereoscope and pathology (Extended Data Fig. 9b) then examined the strength of the 7 GM signatures in each location. This revealed the expected enrichment of GM3 (EMT, IFN, MHC) and GM4 (proliferation) across TNBC cases and GM1 and GM5 (ER, luminal) across ER$^+$ cases (Fig. 6c and Extended Data Fig. 9c). These data suggest that these GMs are not an artifact of dissociation-based methodology. To systematically understand the spatial relationship between modules, we computed Pearson correlations between GM scores in all cancer locations. This revealed two major clusters that were mostly conserved across all six cases, including GM1, GM3, GM5 and GM6 in one cluster, and GM2 and

GM4 in the other (Fig. 6d and Extended Data Fig. 9d). Intriguingly, GM3 (EMT, IFN, MHC) and GM4 (proliferation) showed strong negative correlations in all samples (Fig. 6e–g), suggesting that these distinct cancer phenotypes occur in mutually exclusive regions of breast cancers.

Mapping new heterotypic cellular interactions. While several studies have shown an important role for mesenchymal cells in regulating antitumor immunity[14,55], interactions between stromal and immune cells have yet to be profiled in tissues. Deconvolution revealed spatially distinct subclasses of CAFs, with myofibroblast-like CAFs (CAF s4 and s5) enriched in invasive cancer regions and iCAFs (CAF s1 and s2) dispersed across invasive cancer, stroma and TIL-aggregate regions (Extended Data Fig. 9e). We identified modest negative Pearson correlations between myofibroblast-like CAFs and iCAFs in five of six cases (Fig. 7a–c). Similar CAF localizations were consistent in an independent spatial transcriptomics dataset of 7 HER2$^+$ breast tumors[56], suggesting that this relationship is conserved across clinical subtypes (Fig. 7a). Consistent with the immunoregulatory properties of iCAFs described above, iCAFs colocalized with several lymphocyte populations across both studies, including memory/naive B cells and CD4$^+$/CD8$^+$ T cells (Fig. 7a and Fig. 7d,e). Myofibroblast-like CAFs correlated with CD8$^+$ T cells in six samples (Fig. 7a), suggesting a functional relevance to invasive breast cancers with high TIL infiltration or an immune inflamed phenotype. To explore potential mediators of CAF–lymphocyte interactions at these regions, we investigated the top ligand–receptor interactions at locations most enriched for CAFs and CD4$^+$/CD8$^+$ T cells and that were also detected by these respective cell types by scRNA-seq. This revealed an enrichment of immunoregulatory iCAF ligands and cognate T cell receptors in close proximity, including chemokines (CXCL12/CXCL14-CXCR4 and CXCL10-CXCR3), the complement pathway, transforming growth factor-β (TGFB1/TGFB3-TGFBR2) and lymphocyte inhibitory/activation molecules (LTB-LTBR, TNFSF14-LTBR and LTB-CD40, VTCN1/B7H4-BTLA) (Fig. 7f and Extended Data Fig. 9f). By integrating signaling predictions with cellular proximity, these data highlight relevant candidates for direct regulation of immune cells by CAFs.

Earlier, we defined macrophage states LAM1, LAM2 and Mac:*CXCL10*/c9 with high expression of immunoregulatory molecules such as PD-L1 and PD-L2 (Extended Data Fig. 7g). Across all local Visium cases, LAM1 and LAM2 cells were present at invasive cancer regions; however, LAM2 was also found in areas with high stromal, adipose and lymphocyte cells by morphology (Extended Data Fig. 9e). LAM1 and LAM2 cells showed a modest negative spatial correlation with each other in most cases, which might indicate that a common LAM cell is polarized toward LAM1 or LAM2 by their local TME (Fig. 7a). LAM2 cells, rather than LAM1 cells, were positively correlated with CD4$^+$ and CD8$^+$ T cells in eight tumors across all three subtypes (Fig. 7a). Spots enriched for LAM2 cells and CD4$^+$/CD8$^+$ T cells across multiple tumors coexpressed PD-L1/PD-1 (*CD274*/*PDCD1*) and PD-L2/PD-1 (*PDCD1LG2*/*PDCD1*), suggesting that these cells probably have functional relevance in immunoregulation (Extended Data Fig. 9g). In addition, positive Pearson correlations were identified between Mac:*CXCL10*/c9 and CD8 T cells across many cases (Fig. 7a), which were mostly enriched in spots annotated as invasive cancer + stroma + lymphocytes (Fig. 7g and Extended Data Fig. 9e), suggesting these niches may have functional relevance in regulating antitumor immunity.

Breast tumor ecotypes associated with patient survival. Our single-cell data generated a draft cellular taxonomy of breast tumors, with marked variation and recurring patterns of cellular frequencies observed across 26 tumors. We hypothesized that subsets of breast cancers may have similar cellular composition and tumor biology. To test this at scale, we estimated cellular proportions in large bulk RNA-seq datasets using our single-cell signatures with CIBERSORTx[57]. Estimating cell fractions from pseudobulk samples generated from our single-cell datasets showed good overall correlation between the captured cell fractions and the predicted proportions (median correlation = approximately 0.64), with a majority (32) of cell types showing a significant correlation (Extended Data Fig. 10a). An alternative deconvolution method, DWLS[58], showed similar results (Extended Data Fig. 10b), suggesting that deconvolution methods can effectively predict high-resolution cellular compositions from bulk data.

We deconvoluted all primary breast tumor datasets in the METABRIC cohort[40]. Supporting the validity of the predictions, and SCSubtype, we observed significant enrichment (Wilcoxon test, $P < 2.2 \times 10^{-16}$) of the four SCSubtypes in tumors with matching bulk-PAM50 classifications. Significant enrichment (Wilcoxon test, $P < 2.2 \times 10^{-16}$) of cycling cells in basal, LumB and HER2E tumors was also shown (Extended Data Fig. 10c). Consensus clustering revealed nine tumor clusters with similar estimated cellular composition ('ecotypes') (Fig. 8a–c). These ecotypes displayed correlation with tumor subtype, SCSubtype cell distributions and a diversity of major cell types (Fig. 8a). Ecotype 3 (E3) was enriched for tumors containing Basal_SC, Cycling and Luminal_Progenitor cells (the presumptive cell of origin for basal breast cancers[28]) and a basal bulk PAM50 subtype (Fig. 8a,b). In contrast, E1, E5, E6, E8 and E9 consisted predominantly of luminal cells. Ecotypes also possessed unique patterns of stromal and immune cell enrichment. E4 was highly enriched for immune cells associated with antitumor immunity (Fig. 8a), including exhausted CD8 T cells (*LAG3*/c8), along with T_H1 (*IL7R*/c1) and central memory CD4 T cells (*CCR7*/c0). E2 primarily consisted of LumA and normal-like tumors (Fig. 8b) and was defined by a cluster of mesenchymal cell types, including endothelial CXCL12$^+$ and ACKR1$^+$ cells, s1 MSC iCAFs and depletion of cycling cells (Fig. 8a).

As for prognosis, patients with E2 tumors had the best outcome (Fig. 8d,e), while tumors in E3 were associated with a poor 5-year survival (Fig. 8d), which is consistent with known poor prognosis of basal-like and highly proliferative tumors. E7 also had a poor prognosis and was dominated by HER2E tumors and enrichment of HER2E_SC cells. E4 also had a substantial proportion of HER2E and basal-like tumors (Fig. 8b), yet these patients had significantly better prognosis than E7 (Fig. 8f), perhaps as a consequence of infiltration with antitumor immune cells.

To further assess ecotype robustness, we repeated the consensus clustering using only the 32 significantly correlated cell types and the DWLS method. Substantial overlap of tumors (Supplementary Tables 7 and 8), ecotype features (Extended Data Fig. 10d–f,i,j) and overall survival was seen (Extended Data Fig. 10g,h,k), suggesting that cells with lower deconvolution performance or specific deconvolution methods were not confounding ecotyping.

Finally, we investigated the association between ecotypes and the integrative genomic clusters identified in the METABRIC cohort[40] (Extended Data Fig. 10l). E3 had a high proportion of cancers from integrative genomic cluster 10, which also predominantly consisted of basal-like tumors with similarly poor 5-year survival. E7 had a high proportion of *ERBB2*-amplified and HER2E integrative genomic cluster 5 tumors. These were the worst prognosis groups in both the METABRIC and ecotype analyses. However, most ecotypes did not clearly associate with a specific integrative genomic cluster or PAM50 subtype, which is reflected by the role of stromal and immune cells in defining ecotypes. This lack of unique association suggests that ecotypes are not a simple surrogate for molecular or genomic subtypes.

Discussion

In this study, we provide important advances toward an integrated cellular model for breast cancer classification. We define the cellular architecture of breast tumors at three levels. First, a detailed cellular

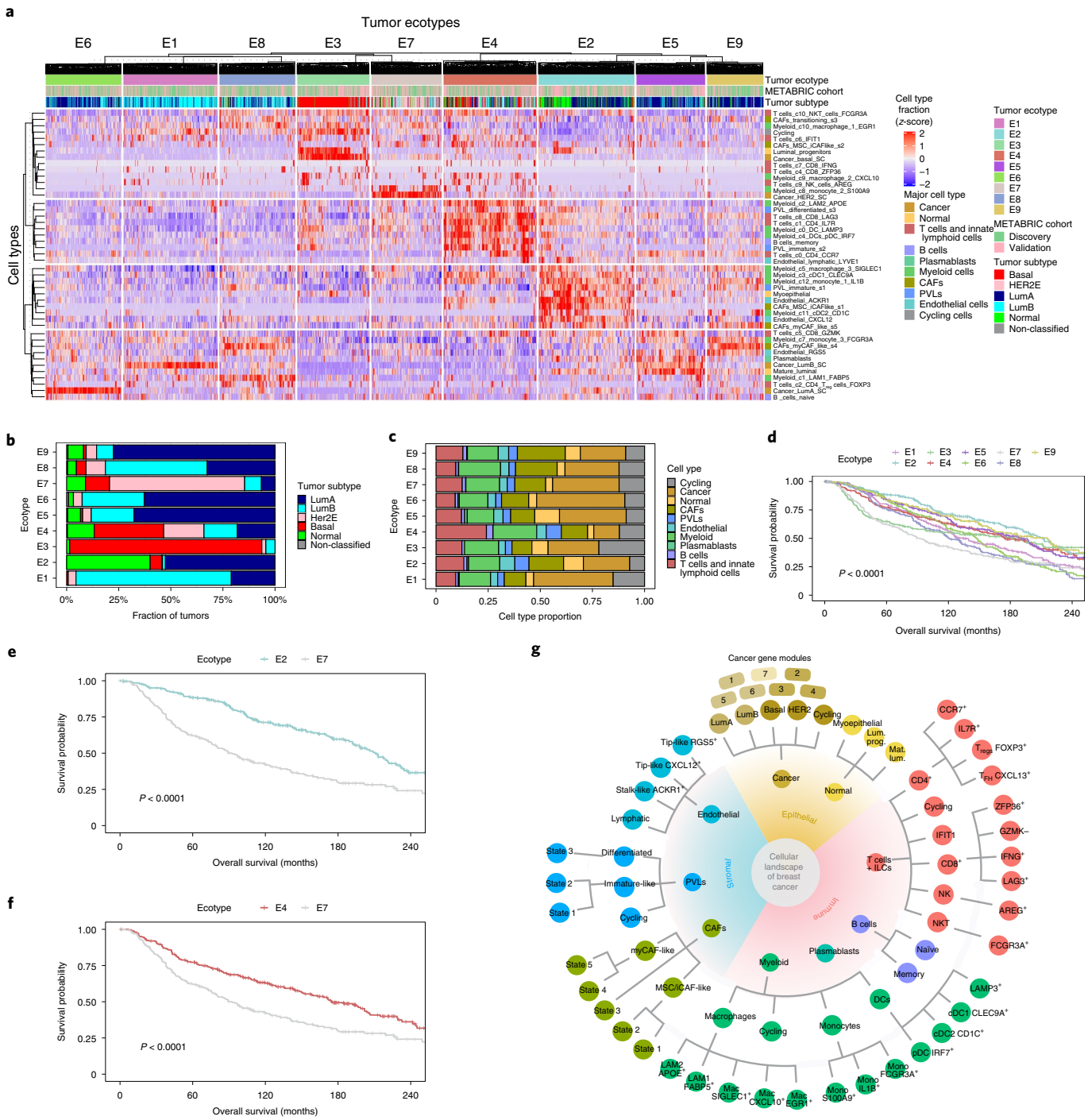


Fig. 8 | Deconvolution of breast cancer cohorts using single-cell signatures reveals robust ecotypes associated with patient survival and intrinsic subtypes. a, Consensus clustering of all tumors (columns) in the METABRIC cohort showing 9 robust tumor ecotypes and 4 groups of cell enrichments from 45 cell types in the breast cancer cell taxonomy. Total 1,985 tumors (E1 = 266, E2 = 269, E3 = 205, E4 = 263, E5 = 195, E6 = 215, E7 = 199, E8 = 213 and E9 = 160). **b**, Relative proportion of the PAM50 molecular subtypes of the tumors in each ecotype. **c**, Relative average proportion of the major cell types enriched in the tumors in each ecotype. **d**–**f**, Kaplan–Meier plot of the patients with tumors in each of the nine ecotypes (**d**), patients with tumors in ecotypes E2 and E7 (**e**) and patients with tumors in ecotypes E4 and E7 (**f**). *P* values were calculated using the log-rank test. **g**, Summary of the epithelial, immune and stromal cell types identified in this study grouped by their major (inner), minor and subset (outer) level classification tiers.

taxonomy that includes new cell types and states and new methods for characterizing cellular heterogeneity (Fig. 8g). Second, a spatial map of cellular locations and interactions within tumors that reveals coordination of tumor and host cell phenotypes within tissue and reveals spatial relationships between cells. Third, using deconvolution, we observed groups of tumors with similar cell type proportions and prognostic associations, named ecotypes, often driven by specific clusters of co-segregating cells.

This study has several limitations. First is the use of tissue dissociation and droplet encapsulation for scRNA-seq, causing certain

cell types including adipocytes, mast cells and granulocytes to be underrepresented. We have addressed this in part by using spatial transcriptomics on intact tissues. Future work may apply complementary technologies, such as single-nucleus or microwell-based sequencing. Second is the limited number of cases per clinical subtype, which limits our ability to estimate subtype-specific features. We used deconvolution to extend our findings into large cohorts of tumors, although these are only estimates of relative cell proportion rather than direct measurements.

Our cellular analysis revealed remarkable heterogeneity for epithelial, immune and mesenchymal phenotypes existing within every tumor, which has confounded previous 'bulk' studies. From this, we derived a high-resolution cellular taxonomy of breast tumors (Fig. 8g) across three tiers of cell types and cell states. We identified at least 9 major cell types that fall into 29 or 49 identifiable states at mid and high resolution, respectively. A number of these states most likely represent dynamic states along a continuum of differentiation, dependent on local interactions. To classify tumor cells in a manner consistent with the previous PAM50 bulk classifier, we developed SCSubtype, which we used to subtype tumors with low cellularity for which bulk analysis had failed. Although heterogeneous expression of subtype markers (for example, cytokeratins, ER) has long been observed in breast cancers, it was not known whether these were simply aberrations in marker expression or reflected functional diversity. SCSubtype provides evidence for the latter, suggesting that intrinsic subtype heterogeneity exists within most cancers. As for all classification methods, the performance of SCSubtype will improve on larger sample sizes applied to the training and test steps in future scRNA-seq studies. Phenotypic diversity in cancer is generally associated with poorer outcomes. While our study is not powered to make this inference, we hypothesize that intratumoral heterogeneity for intrinsic subtype may predict innate resistance to therapy and early relapse after therapy. For instance, the presence of basal-like or HER2-like cells in clinically luminal cancers (Fig. 2c) may cause early relapse after endocrine therapy.

We also conducted an integrative analysis to discover the gene expression programs underlying ITTH. This revealed that GM3 (EMT, IFN, MHC) and GM4 (proliferation) were mutually exclusive, suggesting that a mesenchymal-like state and proliferation are incompatible at cellular resolution. Furthermore, analysis of spatial data revealed organization of these cell states into distinct zones, suggesting a role for the microenvironment in the acquisition of these phenotypes. Proliferation and EMT are inversely correlated in development and previous work in animal models of cancer has shown that exit from a mesenchymal-like state is required for tumor cell proliferation[59]. However, the cellular and spatial relationship between a mesenchymal-like state and proliferation was previously unreported in human cancers. This is particularly interesting in the context of basal-like tumors where both phenotypes predominate, indicating that distinct subsets of cells manifest these phenotypes.

This study has revealed new insights into the immune phenotype of breast tumors. Previous studies have investigated either fewer samples at a similar resolution or a greater number of samples with far fewer parameters[22,23,25,26]. We identified two large clusters of immune cells closely resembling recently identified TREM2-high LAMs[44]. These macrophages also bear similarities to a population of PD-L1^{+} macrophages that associate with high clinical grade and exhausted T cells in breast cancers, identified using mass cytometry[26]. Recent studies have shown that *Trem2*^{hi}-expressing myeloid cells have an immunosuppressive role in mouse models of cancer[11,60], with IHC analyses showing TREM2 expression in multiple subsets of macrophages in TNBC and an association with worse prognosis[60]. Our data extend on these works by providing high-resolution scRNA-Seq, cell-surface protein and spatial characterization of these cells in human cancer. We reveal that LAMs and CXCL10^{hi} macrophages are a major source of immunosuppressive molecules in the human breast TME and spatial analysis revealed their juxtaposition to PD-1^{+} lymphocytes. We also showed that the LAM1 gene signature is associated with poor patient survival in large patient datasets, demonstrating the importance of these cells to breast cancer etiology.

Analysis of the stromal microenvironment revealed three major cell populations—endothelial, CAF and PVL cells—consisting of 3–5 identifiable states each. Previous studies showed that CAF states are interconvertible in distinct tumor culture conditions, suggesting that this differentiation may also occur bidirectionally depending on external factors[17,18]. While differentiation from other progenitors like MSCs is possible, our pseudotemporal analysis provides additional evidence that differentiation can drive transition between CAF subsets. Our observation that mesenchymal subsets are often spatially segregated suggests that signals from the microenvironment control their differentiation or migration. These insights now open pathways to therapeutic strategies aiming to block stromal-immune signaling or manipulate stromal cell differentiation, which may then alter neoplastic and immune cell phenotypes. Importantly, our CITE-seq data provide cell-surface markers for prospective isolation of stromal subsets, enabling ex vivo experimentation.

We used deconvolution to define nine ecotypes among thousands of primary breast cancers. Interestingly, clustering of most ecotypes is driven by cells spanning the major lineages (epithelial, immune and stromal), features not captured by previous studies that stratified disease based on mass cytometry primarily using immune markers[25,26]. Integration of our data with these datasets is an important future direction for the field. While ecotypes are partially associated with intrinsic subtype[4] and genomic classifiers[40], they are not simply surrogates for previous stratification methods. Future work will investigate the molecular mechanisms organizing tissue architecture and tumor ecotypes, aiming to explain their differences in clinical outcome and examine whether tumor ecotypes can be used to personalize therapy.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00911-1.

Received: 11 September 2020; Accepted: 8 July 2021;
Published online: 6 September 2021

References

1. Kim, H. K. et al. Discordance of the PAM50 intrinsic subtypes compared with immunohistochemistry-based surrogate in breast cancer patients: potential implication of genomic alterations of discordance. *Cancer Res. Treat.* **51**, 737–747 (2019).
2. Picornell, A. C. et al. Breast cancer PAM50 signature: correlation and concordance between RNA-Seq and digital multiplexed gene expression technologies in a triple negative breast cancer series. *BMC Genomics* **20**, 452 (2019).
3. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
4. Perou, C. M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
5. Sørlie T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98**, 10869–10874 (2001).
6. Sørlie, T. et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA* **100**, 8418–8423 (2003).
7. Su, S. et al. CD10^{+} GPR77^{+} cancer-associated fibroblasts promote cancer formation and chemoresistance by sustaining cancer stemness. *Cell* **172**, 841–856.e16 (2018).

8. Cazet, A. S. et al. Targeting stromal remodeling and cancer stem cell plasticity overcomes chemoresistance in triple negative breast cancer. *Nat. Commun.* **9**, 2897 (2018).
9. Dushyanthen, S. et al. Relevance of tumor-infiltrating lymphocytes in breast cancer. *BMC Med.* **13**, 202 (2015).
10. Cassetta, L. et al. Human tumor-associated macrophage and monocyte transcriptional landscapes reveal cancer-specific reprogramming, biomarkers, and therapeutic targets. *Cancer Cell* **35**, 588–602.e10 (2019).
11. Katzenelenbogen, Y. et al. Coupled scRNA-seq and intracellular protein activity reveal an immunosuppressive role of TREM2 in cancer. *Cell* **182**, 872–885.e19 (2020).
12. Medler, T. R. et al. Complement C5a fosters squamous carcinogenesis and limits T cell response to chemotherapy. *Cancer Cell* **34**, 561–578.e6 (2018).
13. Nakamura, K. & Smyth, M. J. TREM2 marks tumor-associated macrophages. *Signal Transduct. Target. Ther.* **5**, 233 (2020).
14. Costa, A. et al. Fibroblast heterogeneity and immunosuppressive environment in human breast cancer. *Cancer Cell* **33**, 463–479.e10 (2018).
15. Wu, S. Z. et al. Stromal cell diversity associated with immune evasion in human triple-negative breast cancer. *EMBO J.* **39**, e104063 (2020).
16. Sahai, E. et al. A framework for advancing our understanding of cancer-associated fibroblasts. *Nat. Rev. Cancer* **20**, 174–186 (2020).
17. Öhlund, D. et al. Distinct populations of inflammatory fibroblasts and myofibroblasts in pancreatic cancer. *J. Exp. Med.* **214**, 579–596 (2017).
18. Biffi, G. et al. IL1-induced JAK/STAT signaling is antagonized by TGFβ to shape CAF heterogeneity in pancreatic ductal adenocarcinoma. *Cancer Discov.* **9**, 282–301 (2019).
19. Elyada, E. et al. Cross-species single-cell analysis of pancreatic ductal adenocarcinoma reveals antigen-presenting cancer-associated fibroblasts. *Cancer Discov.* **9**, 1102–1123 (2019).
20. Puram, S. V. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611–1624.e24 (2017).
21. Lambrechts, D. et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat. Med.* **24**, 1277–1289 (2018).
22. Azizi, E. et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**, 1293–1308.e36 (2018).
23. Savas, P. et al. Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nat. Med.* **24**, 986–993 (2018).
24. Kim, C. et al. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell* **173**, 879–893.e13 (2018).
25. Ali, H. R. et al. Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat. Cancer* **1**, 163–175 (2020).
26. Wagner, J. et al. A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell* **177**, 1330–1345.e18 (2019).
27. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).
28. Lim, E. et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in *BRCA1* mutation carriers. *Nat. Med.* **15**, 907–913 (2009).
29. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **16**, 983–986 (2019).
30. Neftel, C. et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* **178**, 835–849.e21 (2019).
31. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
32. Koboldt, D. C. et al. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
33. Prat, A. et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.* **12**, R68 (2010).
34. Nielsen, T. O. et al. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clin. Cancer Res.* **16**, 5222–5232 (2010).
35. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
36. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
37. Glajcar, A., Szpor, J., Hodorowicz-Zaniewska, D., Tyrak, K. E. & Okoń, K. The composition of T cell infiltrates varies in primary invasive breast cancer of different molecular subtypes as well as according to tumor size and nodal status. *Virchows Arch.* **475**, 13–23 (2019).
38. Li, H. et al. Dysfunctional CD8 T cells form a proliferative, dynamically regulated compartment within human melanoma. *Cell* **176**, 775–789.e18 (2019).
39. Yamada, S., Shinozaki, K. & Agematsu, K. Involvement of CD27/CD70 interactions in antigen-specific cytotoxic T-lymphocyte (CTL) activity by perforin-mediated cytotoxicity. *Clin. Exp. Immunol.* **130**, 424–430 (2002).
40. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
41. Slyper, M. et al. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat. Med.* **26**, 792–802 (2020).
42. Ruffell, B. et al. Leukocyte composition of human breast cancer. *Proc. Natl Acad. Sci. USA* **109**, 2796–2801 (2012).
43. Zhang, Q. et al. Landscape and dynamics of single immune cells in hepatocellular carcinoma. *Cell* **179**, 829–845.e20 (2019).
44. Jaitin, D. A. et al. Lipid-associated macrophages control metabolic homeostasis in a Trem2-dependent manner. *Cell* **178**, 686–698.e14 (2019).
45. Chen, J. et al. CCL18 from tumor-associated macrophages promotes breast cancer metastasis via PITPNM3. *Cancer Cell* **19**, 541–555 (2011).
46. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
47. Kumar, A. et al. Specification and diversification of pericytes and smooth muscle cells from mesenchymoangioblasts. *Cell Rep.* **19**, 1902–1916 (2017).
48. Thiriot, A. et al. Differential DARC/ACKR1 expression distinguishes venular from non-venular endothelial cells in murine tissues. *BMC Biol.* **15**, 45 (2017).
49. Mailhos, C. et al. Delta4, an endothelial specific notch ligand expressed at sites of physiological and tumor angiogenesis. *Differentiation* **69**, 135–144 (2001).
50. Ubezio, B. et al. Synchronization of endothelial Dll4-Notch dynamics switch blood vessels from branching to expansion. *eLife* **5**, e12167 (2016).
51. Kryczek, I. et al. CXCL12 and vascular endothelial growth factor synergistically induce neoangiogenesis in human ovarian cancers. *Cancer Res.* **65**, 465–472 (2005).
52. Blanco, R. & Gerhardt, H. VEGF and Notch in tip and stalk cell selection. *Cold Spring Harb. Perspect. Med.* **3**, a006569 (2013).
53. Jakobsson, L. et al. Endothelial cells dynamically compete for the tip cell position during angiogenic sprouting. *Nat. Cell Biol.* **12**, 943–953 (2010).
54. Andersson, A. et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun. Biol.* **3**, 565 (2020).
55. Lakins, M. A., Ghorani, E., Munir, H., Martins, C. P. & Shields, J. D. Cancer-associated fibroblasts induce antigen-specific deletion of $CD8^+$ T cells to protect tumour cells. *Nat. Commun.* **9**, 948 (2018).
56. Andersson, A. et al. Spatial deconvolution of HER2-positive breast tumors reveals novel intercellular relationships. Preprint at *bioRxiv* https://doi.org/10.1101/2020.07.14.200600 (2020).
57. Newman, A. M. et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).
58. Tsoucas, D. et al. Accurate estimation of cell-type composition from gene expression data. *Nat. Commun.* **10**, 2975 (2019).
59. Tsai, J. H., Donaher, J. L., Murphy, D. A., Chau, S. & Yang, J. Spatiotemporal regulation of epithelial-mesenchymal transition is essential for squamous cell carcinoma metastasis. *Cancer Cell* **22**, 725–736 (2012).
60. Molgora, M. et al. TREM2 modulation remodels the tumor myeloid landscape enhancing anti-PD-1 immunotherapy. *Cell* **182**, 886–900.e17 (2020).
61. Leruste, A. et al. Clonally expanded T cells reveal immunogenicity of rhabdoid tumors. *Cancer Cell* **36**, 597–612.e8 (2019).
62. Dutertre, C. A. et al. Single-cell analysis of human mononuclear phagocytes reveals subset-defining markers and identifies circulating inflammatory dendritic cells. *Immunity* **51**, 573–589.e8 (2019).

Methods

Patient material, ethics and consent for publication. The primary untreated breast cancers used in this study (Supplementary Table 1) were collected with written informed consent from all patients under the ×13-0133, ×19-0496, ×16-018 and ×17-155 protocols with approval from all relevant human research ethics committees (Sydney Local Health District Ethics Committee, Royal Prince Alfred Hospital zone and the St Vincent's Hospital Ethics Committee). Consent included the use of all de-identified patient data for publication. Participants were not compensated.

Tissue dissociation. Samples were analyzed from fresh surgical resections and cryopreserved tissue[63]. Tumors were dissociated using the Human Tumor Dissociation Kit (Miltenyi Biotec) according to the manufacturer's protocol. Where viability was <80%, viability enrichment was performed using the EasySep Dead Cell Removal (Annexin V) Kit (STEMCELL Technologies) according to the manufacturer's protocol.

scRNA-seq using 10x Chromium. Single-cell sequencing was performed using the Chromium Single-Cell v2 3′ and 5′ Chemistry Library, Gel Bead, Multiplex and Chip Kits (10x Genomics) according to the manufacturer's protocol. A total of 5,000 to 7,000 cells were targeted per well. Libraries were sequenced on the NextSeq 500 platform (Illumina) with paired-end sequencing and dual indexing. A total of 26, 8 and 98 cycles were run for Read 1, i7 index and Read 2, respectively.

Data processing, cluster annotation and data integration. Raw BCL files were demultiplexed and mapped to the reference genome GRCh38 using the Cell Ranger Single Cell v.2.0 software (10x Genomics). The EmptyDrops method from the DropletUtils package v.1.2.2 (ref. [64]) was applied for cell filtering with additional cutoffs for cells with a gene and unique molecular identifier count greater than 200 and 250, respectively, and a mitochondrial percentage less than 20%. We used the Seurat v.3.0.0 method[36] in R v.3.5.0 for data normalization, dimensionality reduction and clustering using default parameters. Cell clusters were annotated using Garnett[29] v.0.1.4 with classifier-derived breast epithelial cell signatures[28] and immune and stromal cell types from xCell[27]. Data integration was performed using Seurat v.3.0.0 (ref. [36]). (See the Supplementary Note for the specific parameters used.)

Identifying neoplastic from normal breast epithelial cells. The CNV signal for individual cells was estimated using the inferCNV method v.0.99.7 with a 100-gene sliding window. Genes with a mean count of less than 0.1 across all cells were filtered out before the analysis and the signal was denoised using a dynamic threshold of 1.3 s.d. from the mean. Immune and endothelial cells were used to define the reference cell-inferred copy number profiles. Epithelial cells were used for the observations. Epithelial cells were classified into normal (nonneoplastic), neoplastic or unassigned using a similar method to that previously described by Neftel et al.[30]. Briefly, inferred changes at each genomic locus were scaled (between −1 and +1) and the mean of the squares of these values was used to define a genomic instability score for each cell. In each individual tumor, the top 5% of cells with the highest genomic instability scores were used to create an average CNV profile. Each cell was then correlated to this profile. Cells were plotted with respect to both their genomic instability and correlation scores. Partitioning around medoids clustering was performed using the pamk function in the R package cluster v.2.0.7-1 to choose the optimum value for k (between 2 and 4) using silhouette scores and the pam function to apply the clustering. Thresholds defining normal and neoplastic cells were set at 2 cluster s.d. to the left and 1.5 s.d. below the first cancer cluster means. For tumors where partitioning around medoids could not define more than 1 cluster, the thresholds were set at 1 s.d. to the left and 1.25 s.d. below the cluster means. This method was used to identify 27,506 neoplastic and 6,084 normal cells in all tumors; the remaining 3,208 cells were classed as unassigned. Only tumors with at least 200 epithelial cells were used for this neoplastic cell classification step.

Calling PAM50 on pseudobulks and matching bulk RNA-seq. For calling molecular subtypes using the PAM50 method[3], we processed 'pseudobulk' expression profiles for each tumor, named 'Allcells-Pseudobulk', in a similar manner to any bulk RNA-seq sample (that is, upper quartile-normalized, log-transformed). Before PAM50 subtyping, we adjusted a new sample set relative to the PAM50 training set according to their ER and HER2 status as detailed by Zhao et al.[65]. We performed whole-transcriptome RNA-seq using ribosomal depletion (Illumina TruSeq Total RNA) on 24 matching tumor samples from our single-cell dataset. RNA was extracted from diagnostic formalin-fixed paraffin-embedded blocks using the High Pure RNA Paraffin Kit (catalog no. 03 270 289 001; Roche). Libraries were sequenced on the HiSeq 2500 platform (Illumina) with 50-base pair paired-end reads. Transcript quantification was performed using Salmon v0.6.0[66]. We then called PAM50 on each bulk tumor using Zhao et al.[65] normalization and then the PAM50 centroid predictor (Supplementary Table 3).

Calling intrinsic subtype on scRNA-seq using SCSubtype. To design and validate a new subtyping tool specific for scRNA-seq data, we first divided our tumor samples into training and testing sets. The training dataset was defined by identifying tumors with unambiguous molecular subtypes. In this study, we identified robust training set samples using two subtyping approaches: (1) PAM50 subtyping of the Allcells-Pseudobulk datasets (described above); and (2) hierarchical clustering of the Allcells-Pseudobulk data with the 1,100 tumors in the TCGA breast cancer RNA-seq dataset[32] using approximately 2,000 genes from an intrinsic breast cancer gene list[3]. We first identified tumors that shared the same 'concordant' subtype from both Allcells-Pseudobulk PAM50 calls and TCGA hierarchical clustering-based subtype classifications (Supplementary Table 3). Next, since our methodology aimed to subtype cancer cells, we removed any tumors with <150 cancer cells. Finally, we did not include cells from the two metaplastic samples (CID4513 and CID4523) in the training data because this was a histological subtype not used in the original PAM50 training set. Only tumor cells with >500 unique molecular identifiers were used for the training and test datasets in SCSubtype (total of 24,889 cells). Within each subtype training set, we utilized the cancer cells from each tumor sample and performed pairwise single-cell integrations and differential gene expression calculations. The integration was carried out in a 'within-group' pairwise fashion using the FindIntegrationAnchors and IntegrateData functions in Seurat v.3.0.0 (ref. [36]). Briefly, the first step identifies anchors between pairs of cells from each dataset using mutual nearest neighbors. The second step integrates the datasets together based on a distance-based weight matrix constructed from the anchor pairs. Differentially expressed genes were calculated between each pair using a Wilcoxon rank-sum test by the FindAllMarkers function within Seurat. The following pairs were analyzed: HER2E (CID3921-CID44991, CID44991-CID45171, CID45171-CID3921); basal-like (CID4495-CID44971, CID44971-CID4515, CID4515-CID4495); LumA (CID4290-CID4530); and LumB (CID3948-CID4535). We removed any duplicate genes occurring between the 4 training groups, which yielded 4 sets of genes composed of 89 genes defining Basal_SC, 102 genes defining HER2E_SC, 46 genes defining LumA_SC and 65 genes defining LumB_SC, which we defined as SCSubtype gene signatures (Supplementary Table 4). To assign a subtype call to a cell, we calculated the average (that is, the mean) read counts for each of the four signatures for each cell. The SC subtype with the highest signature score was then assigned to each cell. We utilized this method to subtype all 24,489 neoplastic cells, from both our training samples ($n=10$) and the remaining test ($n=10$) set samples. '

Calculating proliferation and differentiation scores. We calculated the degree of epithelial cell differentiation (DScore)[33] and proliferation[34] on all tumor cells from our scRNA-seq cohort and 1,100 tumors from the TCGA dataset. The DScore was computed using a centroid-based predictor with information from approximately 20,000 genes[33]. Averaged normalized expression of 11 genes[34] (*BIRC5*, *CCNB1*, *CDC20*, *NUF2*, *CEP55*, *NDC80*, *MKI67*, *PTTG1*, *RRM2*, *TYMS* and *UBE2C*), independent of the SCSubtype gene lists, was used to compute the proliferation score.

Histology and immunohistochemical staining of CK5 and ER. Tumor tissue was fixed in 10% neutral buffered formalin for 24 h and then processed for paraffin embedding. Diagnostic tumor blocks were accessed for samples that did not have a research block available. Blocks were sectioned at 4 µM. Sections were stained with hematoxylin and eosin (H&E) for standard histological analysis. IHC was performed on serial sections with prediluted primary antibodies against ER (clone 6F11, catalog no. PA0151; Leica Biosystems) or CK5 (clone XM26, catalog no. PA0468; Leica Biosystems) using suggested protocols on the BOND RX Autostainer (Leica Biosystems). Antigen retrieval was performed for 20 min using the BOND Epitope Retrieval solution 1 for ER or solution 2 for CK5, followed by primary antibody incubation for 60 min and secondary staining with the Bond Refine Detection System (Leica Biosystems). Slides were imaged using the Aperio CS2 Digital Pathology Slide Scanner and processed with QuPath v.0.2.0.

GM analysis of neoplastic intratumor heterogeneity. For each individual tumor, with more than 50 neoplastic cells, neoplastic cells were clustered using Seurat v.3.0.0 (ref. [36]) at five resolutions (0.4, 0.8, 1.2, 1.6, 2.0). MAST[67] v.1.12.0 was then used to identify the top 200 differentially regulated genes in each cluster. Only gene signatures containing more than five genes and originating from clusters of more than five cells were kept. In addition, redundancy was reduced by comparing all pairs of signatures within each sample and removing the pair with the fewest genes from those pairs with a Jaccard index >0.75. Across all tumors, a total of 574 gene signatures of intratumor heterogeneity were identified.

Consensus clustering (using spherical k-means, SKmeans, implemented in the cola R package v.1.2.0 (https://www.bioconductor.org/packages/release/bioc/html/cola.html)) of the Jaccard similarities between these gene signatures was used to identify seven robust groups or GMs. For each of these, a GM was defined by taking the 200 genes that had the highest frequency of occurrence across clusters and individual tumors. These are defined as GM1 to GM7. A GM signature was calculated for each cell using AUCell v1.4.1[68] and each neoplastic cell was assigned to a module, using the maximum of the scaled AUCell GM signature scores. This resulted in 4,368, 3,288, 2,951, 4,326, 3,931, 2,500 and 3,125 cells assigned to GM1 to GM7, respectively. These are defined as GM-based neoplastic cell states.

Differential gene expression, module and pathway enrichment. Differential gene expression was performed using MAST[67] v.1.8.2. All DEGs from each cluster (log fold change >0.5, *P* threshold of 0.05 and adjusted *P* threshold of 0.05; Supplementary Tables 9 and 10) were used as input into the ClusterProfiler package[69] v.3.14.0 for gene ontology functional enrichment. Results were clustered, scaled and visualized using the pheatmap package v.1.0.12. Cytotoxic, TAM and dysfunctional T cell gene expression signatures were assigned using the AddModuleScore function in Seurat v.3.0.0 (ref. [36]). The list of genes used for dysfunctional T cells were adopted from Li et al.[38]. The TAM gene list was adopted from Cassetta et al.[10]. The cytotoxic gene list consists of 12 genes that translate to effector cytotoxic proteins (*GZMA*, *GZMB*, *GZMH*, *GZMK*, *GZMM*, *GNLY*, *PRF1* and *FASLG*) and well-described cytotoxic T cell activation markers (*IFNG*, *TNF*, *IL2R* and *IL2*).

Pseudotemporal ordering to infer cell trajectories. Cell differentiation was inferred for mesenchymal cells (CAFs, PVLs and endothelial cells) using the Monocle 2 (ref. [46]) v.2.10.1 with default parameters as recommended by the developers. Integrated gene expression matrices from each cell type were first exported from Seurat v.3 into Monocle to construct a CellDataSet. All variable genes defined by the differentialGeneTest function (cutoff of $q < 0.001$) were used for cell ordering with the setOrderingFilter function. Dimensionality reduction was performed with no normalization methods and the DDRTree reduction method in the reduceDimension step.

CITE-seq antibody staining. Samples were stained with 10x Chromium 3′ messenger RNA capture compatible TotalSeq-A antibodies (BioLegend). A total of four cases from our scRNA-seq cohort were analyzed with a panel of 157 barcoded antibodies (Supplementary Table 11), including one luminal (CID4040), one HER2 (CID3838) and two TNBC (CID4515 and CID3956). Staining was performed as described previously by Stoeckius et. al.[35]. Briefly, a maximum of 1 million cells per sample was resuspended in 120 µl of cell staining buffer (BioLegend) with 5 µl of Fc Receptor Block (TruStain FcX; BioLegend) for 15 min. This was followed by a 30-min staining of the antibodies at 4 °C. A concentration of 1 µg 100 µl^{-1} was used for all antibody markers used in this study. The cells were then washed 3 times with PBS containing 10% FCS medium followed by centrifugation (300 *g* for 5 min at 4 °C) and expungement of supernatant.

CITE-seq data processing and imputation. Demultiplexed reads were assigned to individual cells and antibodies with the Python package CITE-seq-Count v.1.4.3 (https://github.com/Hoohm/CITE-seq-Count/tree/1.4.2). CITE counts were normalized and scaled with Seurat v.3.1.4. Imputation of CITE data was performed per individual cell type (B, T, myeloid and mesenchymal cells) for those antibodies that were differentially expressed between subclusters (FindAllMarkers step) for individual samples. We used anchoring-based transfer learning to transfer protein expression levels from these four samples to the remaining cases[36].

Spatial transcriptomics. Tissue samples were embedded in optimal cutting temperature compound and stored at −80 °C. Tissue blocks were cut into 10-µm sections and processed using the Visium Spatial Gene Expression Kit (10x Genomics) according to the manufacturer's instructions. First, breast tissue permeabilization condition was optimized using the Visium Spatial Tissue Optimization Kit, which was found to be ideal at 12 min. Sections were stained with H&E and imaged using a Leica DM6000 microscope under a 20× lens magnification, then processed for spatial transcriptomics. The resulting complementary DNA library was checked for quality control, then sequenced using an Illumina NovaSeq 6000 system. Cycling conditions were set for 28, 98 and 8 for Read 1, Read 2 and Read 3 (i7 index), respectively. Spots were annotated by a specialist breast pathologist using the Loupe v.4.0.0 software (10x Genomics).

Visium spatial transcriptomics data processing. Reads were demultiplexed and mapped to the reference genome GRCh38 using the Space Ranger software v.1.0.0 (10x Genomics). Count matrices were loaded into Seurat v.3.2.0 and STutility v.0.1.0 for all subsequent data filtering, normalization, filtering, dimensional reduction and visualization. Data normalization was performed on independent tissue sections using the variance-stabilizing transformation method implemented in the SCTransform function in Seurat. We applied nonnegative matrix factorization to the normalized expression matrix using STutility (nFactors = 20).

Spatial deconvolution using Stereoscope. We performed deconvolution of spatial tissue locations using Stereoscope[54] v.0.2.0, a probabilistic model for estimating cell type proportions using annotated scRNA-seq data as input. Stereoscope was performed using default parameters (Supplementary Note). We matched spatial and single-cell data with respect to breast cancer clinical subtype. We deconvolved cell types across three tiers of classification including major, minor and subset lineages.

Mapping cancer heterogeneity and cell signaling predictions. To investigate breast cancer GMs, we first filtered all spots where cancer epithelial cells were called using the Stereoscope method with a filter of 10%. GM gene lists were then scored using AUCell[68] v.1.4.1. GM correlations were then computed using Pearson correlation across all spots in R (cor.test function; cutoff $P = 0.05$). For cell–cell colocalizations across all tissue domains, we included seven additional HER2$^+$ datasets generated on a platform similar to Visium[56]. In total, Pearson correlation was computed from the cell abundances across the tissue locations from 13 patients using *R* (cor.test function; cutoff $P = 0.05$). For cell signaling predictions between iCAFs and CD4/CD8$^+$ T cells, spots containing the two cell types of interest were first selected using the product of the two respective deconvolution values. Interaction scores were defined as the product of the ligand and receptor log expression levels using two independent cell signaling sets[70,71] and only ligands and receptors differentially expressed by iCAFs and CD4/CD8$^+$ T cells in the scRNA-seq data, respectively (MAST; average log fold change threshold = 0.1). All regions annotated as normal ductal by pathology were also excluded from the above analyses.

Survival analysis of scRNA-seq signatures. To assess the impact of particular cell types described by scRNA-seq (for example, LAM1 and LAM2) on clinical outcome, we assessed the association between gene signatures (derived as described above) with patient overall survival in the METABRIC cohort. For each tumor from the bulk expression cohort, average gene signature expression was derived using the top 100 genes from the gene signature of interest. Patients were then stratified based on the top and bottom 30%; survival curves were generated using the Kaplan–Meier method with the 'survival' package v.2.44-1.1. We assessed the significance between two groups using the log-rank test statistics.

Tumor ecotype analysis using deconvolution. CIBERSORTx[57] v.1.0 and DWLS[58] (accessed from https://github.com/dtsoucas/DWLS on 30/11/2020) were used to deconvolute predicted cell fractions from a number of bulk transcript profiling datasets (see Supplementary Note for specific parameters). To prevent confounding of cycling cell types, we first assigned all neoplastic epithelial cells with a proliferation score greater than 0 as cycling and then combined these with cycling cell states from all other cell types to generate a single cycling cell state. Normalized METABRIC expression matrices, clinical information and PAM50 subtype classifications were obtained from METABRIC (https://www.cbioportal.org/study/summary?id=brca_metabric). Tumor ecotypes in the METABRIC cohort were identified using SKmeans-based consensus clustering (as implemented in cola v.1.2.0) of the predicted cell fraction from either CIBERSORTx or DWLS in each bulk METABRIC patient tumor. When comparing ecotypes between methods (that is, consensus clustering results from using the cell abundances of all cell types or just the 32 significantly correlated cell types from CIBERSORTx deconvolution and the consensus clustering results from CIBERSORTx or DWLS cell abundances), the number of tumor ecotypes was fixed as 9 and the tumor overlaps between all ecotype pairs was calculated (Supplementary Tables 7 and 8). Common ecotypes were then identified by identifying the ecotype pairs with the largest average METABRIC tumor overlap. Differences in survival between ecotypes were assessed using Kaplan–Meier analysis and log-rank test statistics using the survival v.2.44-1.1 and survminer v.0.4.7 R packages.

Statistics and reproducibility. No statistical method was used to predetermine sample size. Statistical significance for DEGs were determined using the Wilcoxon rank-sum test, with all *P* values adjusted using Bonferroni correction. All box plots depict the first and third quartiles as the lower and upper bounds, respectively. The whiskers represent 1.5× the interquartile range (IQR) and the center depicts the median. All statistical tests used are defined in the figure legends.

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

Data availability

All processed scRNA-seq data are available for in-browser exploration and download through the Broad Institute Single Cell portal at https://singlecell.broadinstitute.org/single_cell/study/SCP1039. Processed scRNA-seq data from this study are also available through the Gene Expression Omnibus under accession number GSE176078. Raw scRNA-seq data from this study have been deposited with the European Genome-phenome Archive, which is hosted by the European Bioinformatics Institute and Centre for Genomic Regulation under accession no. EGAS00001005173. All spatially resolved transcriptomics data from this study are available from the Zenodo data repository (https://doi.org/10.5281/zenodo.4739739). Spatially resolved transcriptomics data from Andersson et al.[56] can be downloaded from the Zenodo data repository (https://doi.org/10.5281/zenodo.3957257).

Code availability

Code related to the analyses in this study can be found on GitHub at https://github.com/Swarbricklab-code/BrCa_cell_atlas (ref. [72]).

References

63. Wu, S. Z. et al. Cryopreservation of human cancers conserves tumour heterogeneity for single-cell multi-omics analysis. *Genome Med.* **13**, 81 (2021).
64. Lun, A. T. L. et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).
65. Zhao, X., Rødland, E. A., Tibshirani, R. & Plevritis, S. Molecular subtyping for clinically defined breast cancer subgroups. *Breast Cancer Res.* **17**, 29 (2015).
66. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
67. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
68. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
69. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
70. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).
71. Ramilowski, J. A. et al. A draft network of ligand–receptor-mediated multicellular signalling in human. *Nat. Commun.* **6**, 7866 (2015).
72. Swarbrick, A., Wu, S., Al-Eryani, G., Roden, D. & Bartonicek, N. BrCa_cell_atlas. Version 1.0.0 (analysis code) https://doi.org/10.5281/zenodo.5031502 (2021).

Acknowledgements

This work is supported by a research grant from the National Breast Cancer Foundation (NBCF) of Australia (no. IIRS-19-106) and supported by the generosity of J. McMurtrie, AM and D. McMurtrie, the Petre Foundation, White Butterfly Foundation, Sydney Breast Cancer Foundation, Skipper Jacobs Charitable Trust, G. P. Harris Foundation and The National Health and Medical Research Council (NHMRC). A.S. is the recipient of a Senior Research Fellowship from the NHMRC (no. APP1161216). S.Z.W., G.A.-E. and J.T. are supported by the Australian Government Research Training Program Scholarship. S.O.T. is supported by the NBCF (PRAC 16-006; no. IIRS-19-084), Sydney Breast Cancer Foundation and the Family and Friends of M. O'Sullivan. S.J. is supported by a research fellowship from the NBCF. X.S.L. is supported by the Breast Cancer Research Foundation (no. BCRF-19-100) and National Institutes of Health (no. R01CA234018). C.M.P. and A.T. were supported by the National Cancer Institute Breast SPORE program (no. P50-CA58223), grant no. RO1-CA148761, and Breast Cancer Research Foundation. This work was supported by the Australian Centre for Translational Breast Cancer Research, Walter and Eliza Hall Institute, with funding support from the NHMRC Centre for Research Excellence grant no. APP1153049. E.L. is supported as a National Breast Cancer Foundation Endowed Chair and by the Love Your Sister foundation. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank the following people for their assistance in the experimental part of this manuscript: J. Yang; G. Lehrbach from the Garvan Institute of Medical Research Tissue Culture Facility; A. Zaratzian from the Garvan Histopathology Facility for tissue processing and IHC staining and guidance on the Visium experiments; the Garvan–Weizmann Centre for Cellular Genomics, including E. Lam, H. Saeed and M. Armstrong for the expertise in flow sorting. We thank H. Holliday for the incredible illustration in Fig. 8g. We thank H. H. Milioli for providing guidance for analyzing the METABRIC cohort dataset. We thank I. Shapiro and C. Grant as consumer advocates. This manuscript was edited at Life Science Editors.

Author contributions

A.S. conceived the project and directed the study with input from all authors. E.L., S.W., M.N.H., B.C., C.C., C.M., D.S., E.R., A.P., J.B., S.O.T., E.M. and L.G. contributed to the experimental design, procured the patient tumor tissue and assisted with interpreting the data. S.Z.W., G.A.-E. and K.H. performed the single-cell captures. K.H. analyzed all the clinical information. S.Z.W., K.H. and G.A.-E. optimized and performed the tumor dissociation experiments. G.A.-E. optimized and performed the antibody staining for the CITE-seq experiments. N.B. and G.A.-E. performed the CITE-seq data processing. C.-L.C. and S.Z.W. performed the scRNA-seq experiments on the Chromium Controller. C.-L.C. helped perform the next-generation sequencing of the scRNA-seq libraries. S.Z.W. performed the preprocessing, data integration and reclustering steps for the scRNA-seq data. J.T. performed the analysis and benchmarking of inferCNV. A.T. and C.M.P. led the development of SCSubtype. D.R. interpreted and led the analyses for the breast cancer GM analyses. K.H. and T.W. performed the H&E and IHC experiments. S.O.T. independently assessed and scored all histology in this study. G.A.-E. interpreted and performed the analyses of the immune cells with intellectual input from S.J. C.-A.D. and F.G. provided intellectual input related to myeloid cluster annotation. S.Z.W. interpreted and performed all the analyses of stromal cells. D.K. and C.-L.C. performed the Visium experiments with input from J.E.P. V.G. helped perform preprocessing of the Visium datasets. S.R.W., N.I.W., C.R.U., J.G.C. and Z.W.B. performed the Visium experiments and data processing from an independent laboratory. A.A. performed the Stereoscope deconvolution with input from J.L. S.Z.W. performed the downstream analysis of the Visium data with guidance from A.A., L.L., G.A.-E. and J.L. D.R. interpreted and performed the CIBERSORTx analysis. S.Z.W. and D.R. performed the survival analyses. C.W. and X.S.L. provided intellectual input and guidance on bulk deconvolution and survival analyses. S.Z.W., A.S., D.R., G.A.-E. and S.J. wrote the manuscript with input from all authors.

Competing interests

C.M.P. is an equity stockholder and consultant for BioClassifier; he is also listed as an inventor on patent applications for the Breast PAM50 Subtyping assay. J.L. is an author on patents owned by Spatial Transcriptomics AB covering technology presented in this paper. The other authors declare no competing interests.

Additional information

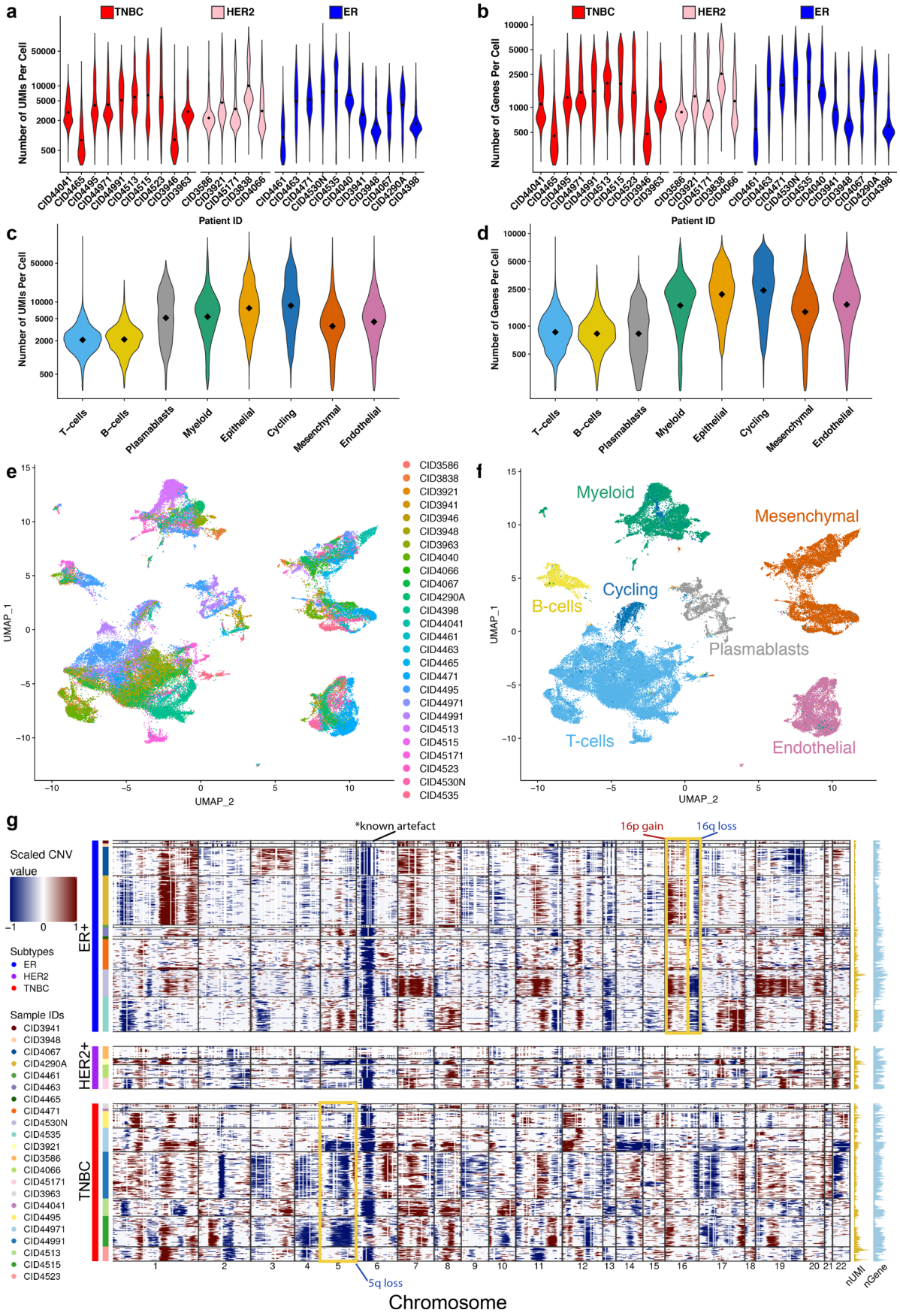
Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00911-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00911-1.

Correspondence and requests for materials should be addressed to Alexander Swarbrick.

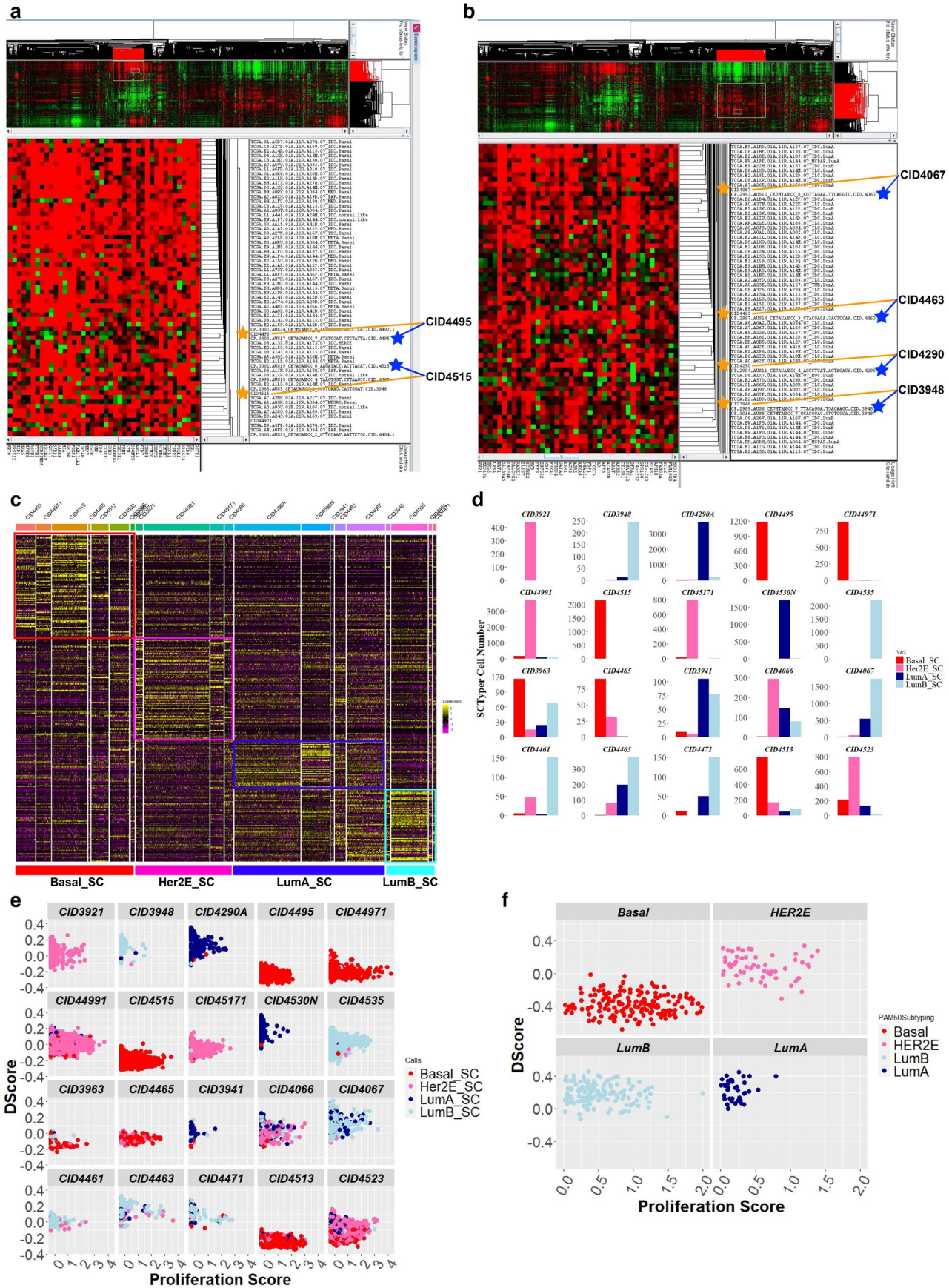
Peer review information *Nature Genetics* thanks Itai Yanai and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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



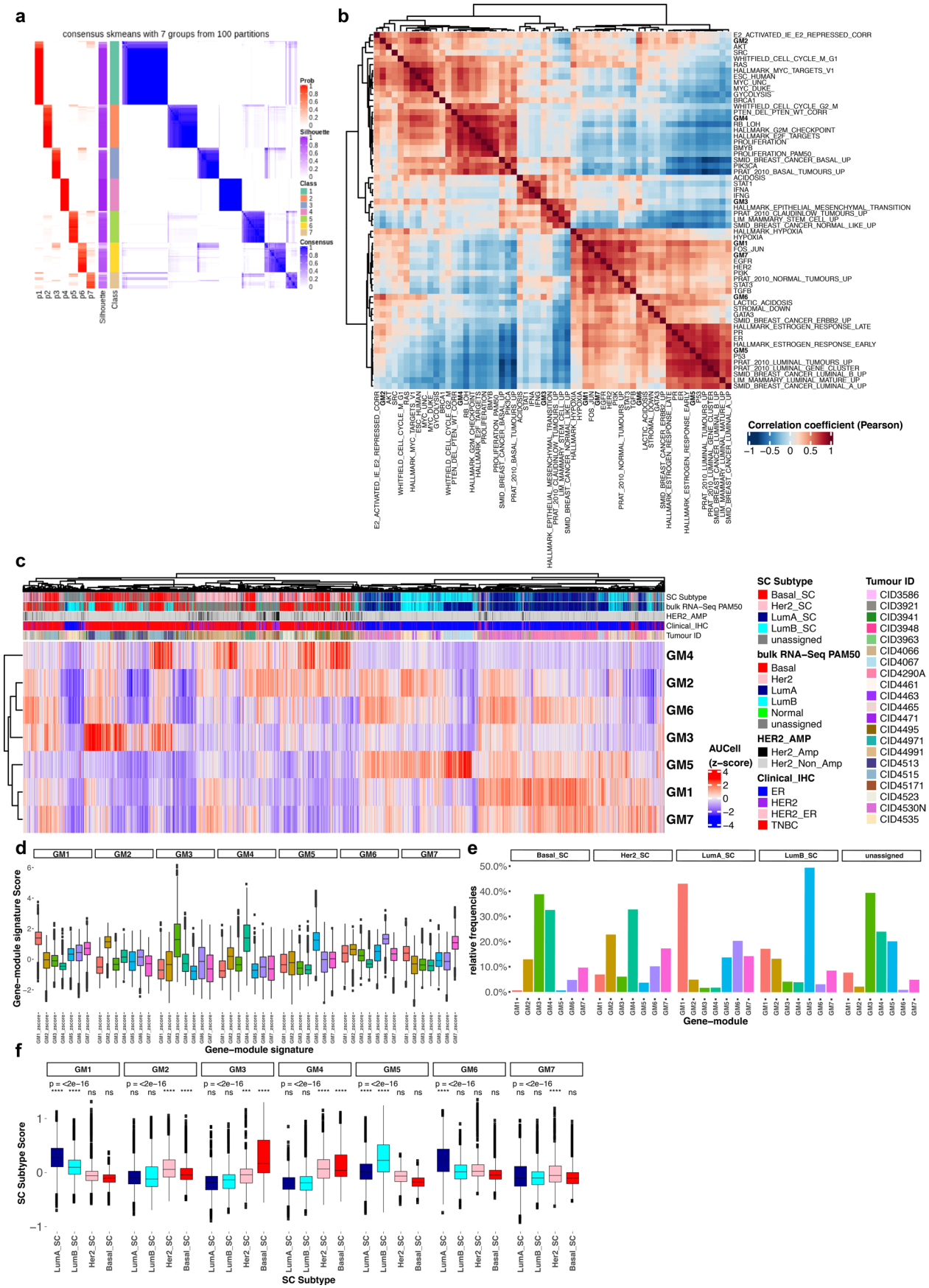
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Identification of malignant cells, single-cell RNA sequencing metrics and non-integrated data of stromal and immune cells. **a-b**, Number of unique molecular identifiers (**a**) and genes (**b**) per tumor analyzed by scRNA-Seq in this study. Tumors are stratified by the clinical subtypes TNBC (red), HER2 (pink) and ER (blue). Diamond points represent the mean. **c-d**, Number of unique molecular identifiers (UMIs;**c**) and genes (**d**) per major lineage cell types identified in this study. These major lineage tiers are grouped by T-cells, B-cells, Plasmablasts, Myeloid, Epithelial, Cycling, Mesenchymal (cancer-associated fibroblasts and perivascular-like cells) and Endothelial. Diamond points represent the mean. **e-f**, UMAP visualization of all 71,220 stromal and immune cells without batch correction and data integration. UMAP dimensional reduction was performed using 100 principal components in the Seurat v3 package. Cells are grouped by tumor (**e**) and major lineage tiers (**f**) as identified using the Garnett cell classification method. **g**, InferCNV heatmaps of all malignant cells grouped by clinical subtypes. Common subtype-specific CNVs and a chr6 artefact reported by Tirosh et. al. are marked (Tirosh et al., 2016b).



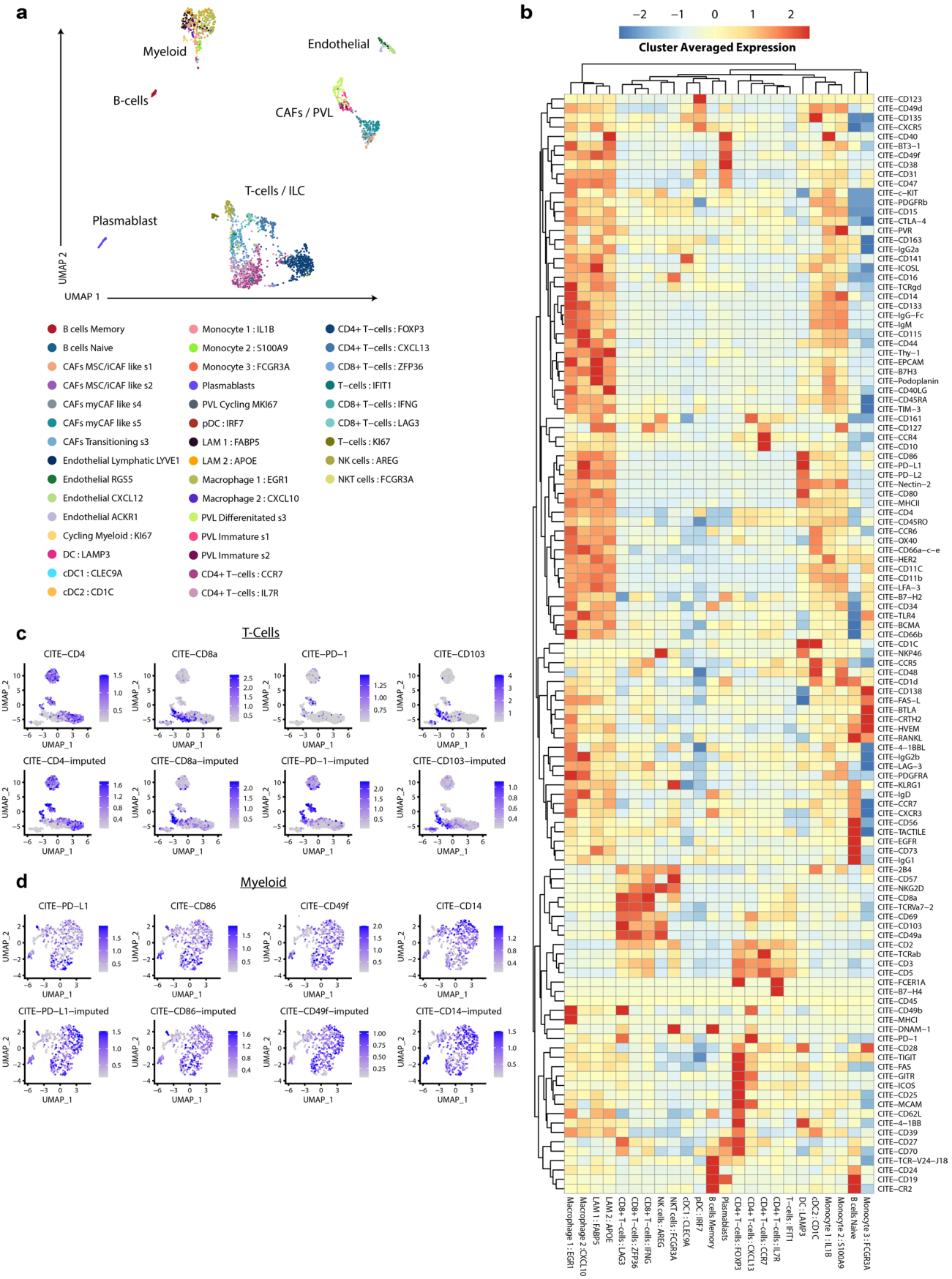
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Supplementary data for SCSubtype classifier. a-b, Hierarchical Clustering of Allcells-Pseudobulk (indicated by yellow stars) and Ribozero mRNA-Seq (indicated by blue stars) profiles of the patient samples with TCGA patient mRNA-Seq data. **a**, View of the basal cluster showing pairing of Allcells-Pseudobulk and Ribozero mRNA-Seq profiles of 2 representative tumors (CID4495 and CID4515) in the present study. **b**, View of the luminal cluster showing pairing of Allcells-Pseudobulk and Ribozero mRNA-Seq profiles of 4 representative tumors (CID4067, CID4463, CID4290 and CID3948) in the present study. **c**, Heatmap of SCSubtype gene sets across the training and test samples in each individual group. Colored outlined boxes highlighting the top expressed genes per group. **d**, Barplot representing proportions of SCSubtype calls in individual samples. Test dataset samples are highlighted within the golden colored outline. **e**, Scatterplot of individual cancer cells plotted according to the Proliferation score (x-axis) and Differentiation – DScore (y-axis). Individual cells are colored based on the SCSubtype calls. **f**, Scatterplot of individual TCGA breast tumors plotted according to the Proliferation score (x-axis) and Differentiation – DScore (y-axis). Individual patients are colored based on the PAM50 subtype calls.



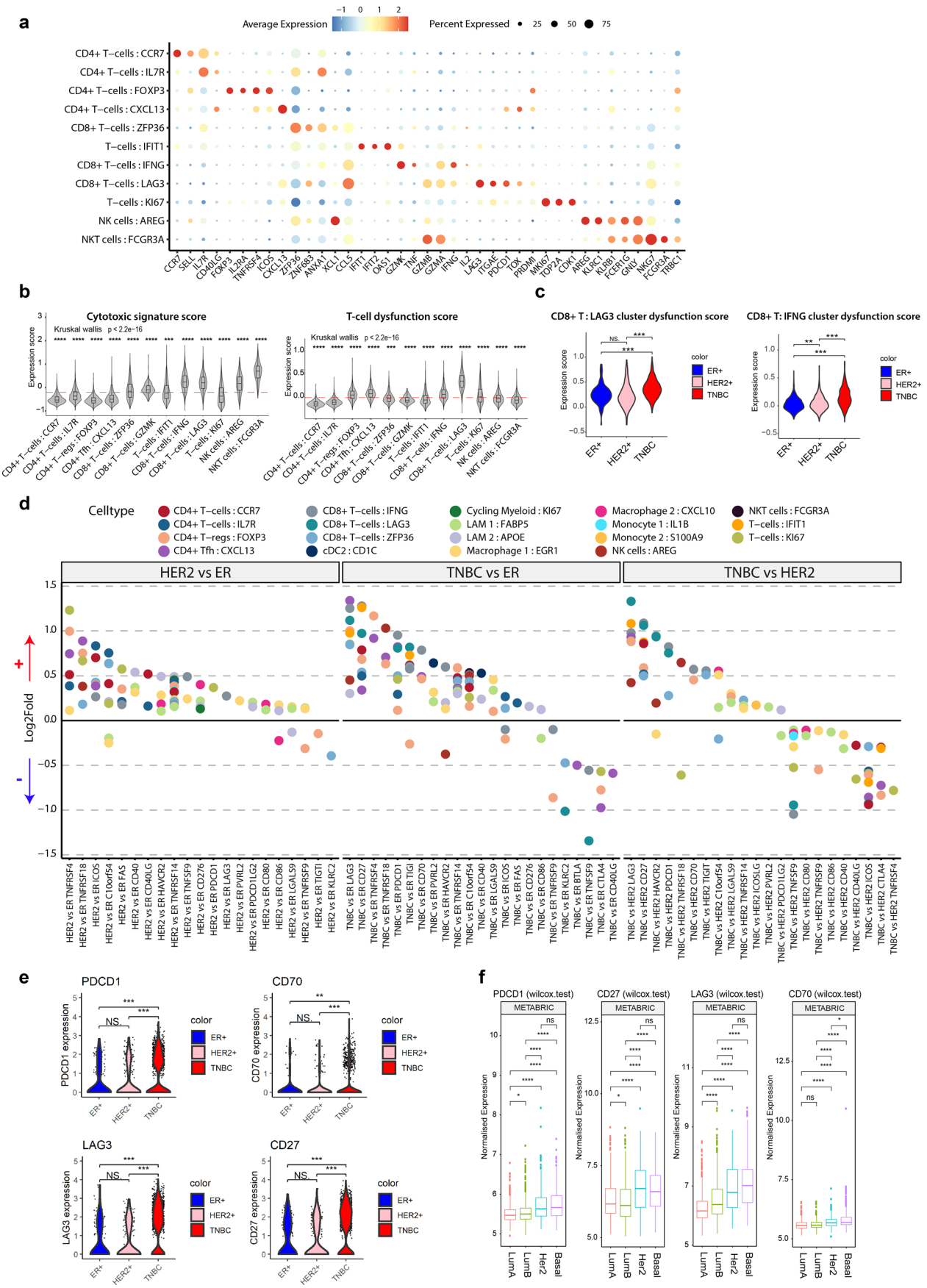
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Supplementary data for breast cancer gene modules. a, Spherical k-means (skmeans) based consensus clustering of the Jaccard similarities between 574 signatures of neoplastic cell ITTH. This showed the probability (p1-p7) of each signature of ITTH being assigned to one of seven clusters/classes. Silhouette scores are shown for each signature. **b**, Heatmap of pair-wise Pearson correlations of the scaled AUCell signature scores, across all individual neoplastic cells, for each of the seven ITTH gene-modules (bolded) and a curated set of breast cancer related gene-signatures. Hierarchical clustering was performed using Pearson correlations and average linkage **c**, Heatmap showing the scaled AUCell signature scores of each of the seven ITTH gene-modules (rows) across all individual neoplastic cells (columns). Hierarchical clustering was done using Pearson correlations and average linkage. (HER2_AMP = Clinical HER2 amplification status). **d**, Distributions of signature scores (z-score scaled) for each of the gene-module signatures (24,489 cells from 21 tumors). Cells are grouped according to the gene-module (GM1-7) cell-state. **e**, Barchart showing the proportion of cells assigned to each of the gene-module cell-states (GM1-7) with cells grouped according to the SCSubtypes. **f**, Distributions of SCSubtype scores for each of the gene-module signatures (24,489 cells from 21 tumors). Cells are grouped according to the gene-module (GM1-7) cell-state. Kruskal-Wallis tests were performed to calculate the significance between the four SCSubtype score groups in each of the gene-module groups, p-value shown. Wilcox tests were used to identify which SCSubtype had significantly increased SCSubtype scores in the cells assigned to each gene-module, the scores of each SCSubtype were compared to the rest of the SCSubtype scores (****: Holm adjusted p-value < 0.0001, ns: Holm adjusted p-value > 0.05). Box plots in d and f depict the first and third quartiles as the lower and upper bounds, respectively. The whiskers represent 1.5x the interquartile range and the centre depicts the median.



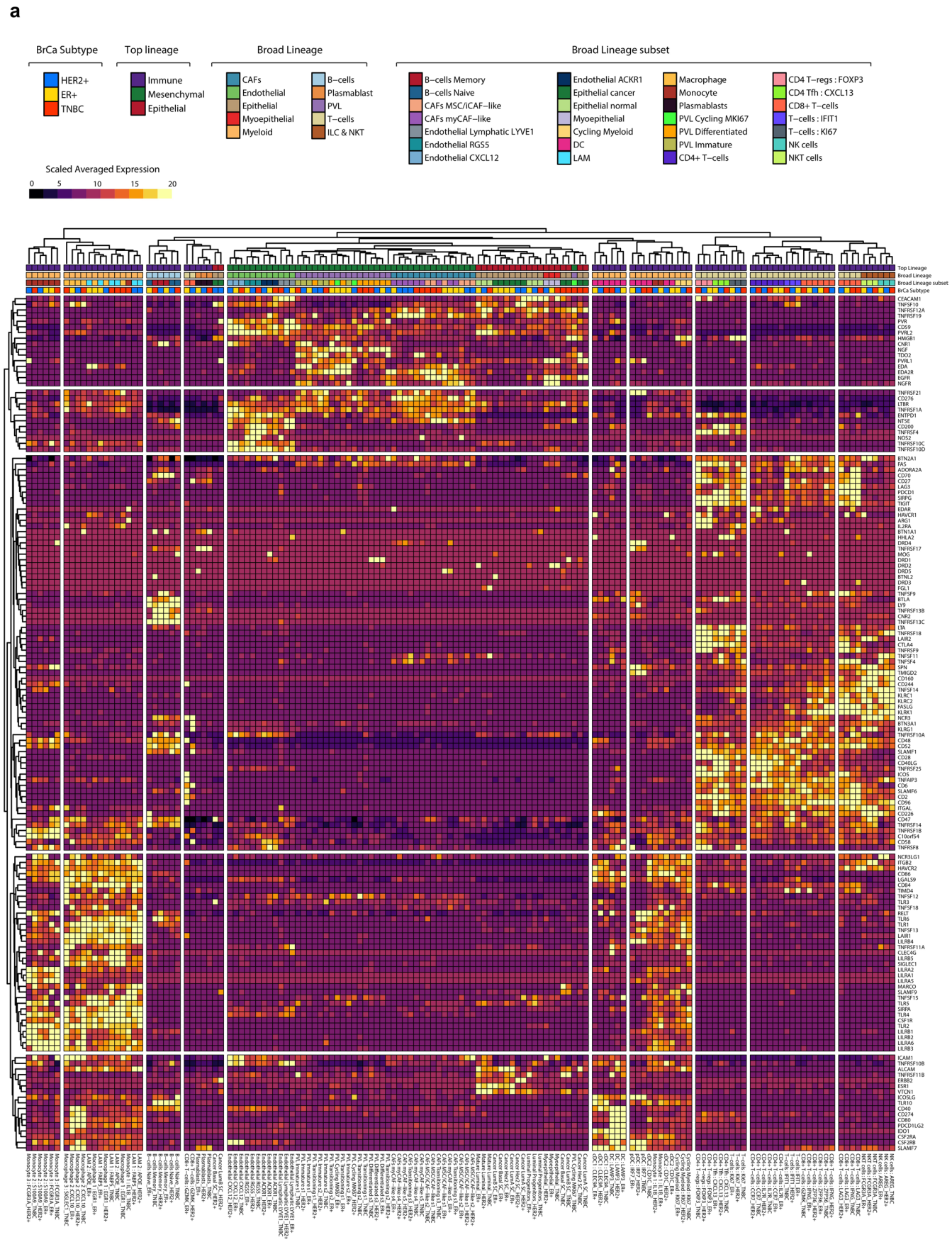
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | CITE-seq vignette. a, UMAP Visualization of a TNBC sample with 157 DNA barcoded antibodies (Supplementary Table 11). Cluster annotations were extracted from our final breast cancer atlas cell annotations. **b**, Heatmap visualization of the cluster averaged antibody derived tag (ADT) values for the 157 CITE-seq antibody panel. Only immune cells are shown. **c-d**, Expression featureplots of measured experimental ADT values (shown in top rows) against the CITE-seq imputation ADT levels (shown in bottom rows), as determined using the seurat v3 method. Selected markers for immunophenotyping T-cells (**c**; CD4, CD8A, PD-1 and CD103) and myeloid cells (**d**; PD-L1, CD86, CD49f and CD14) are shown.



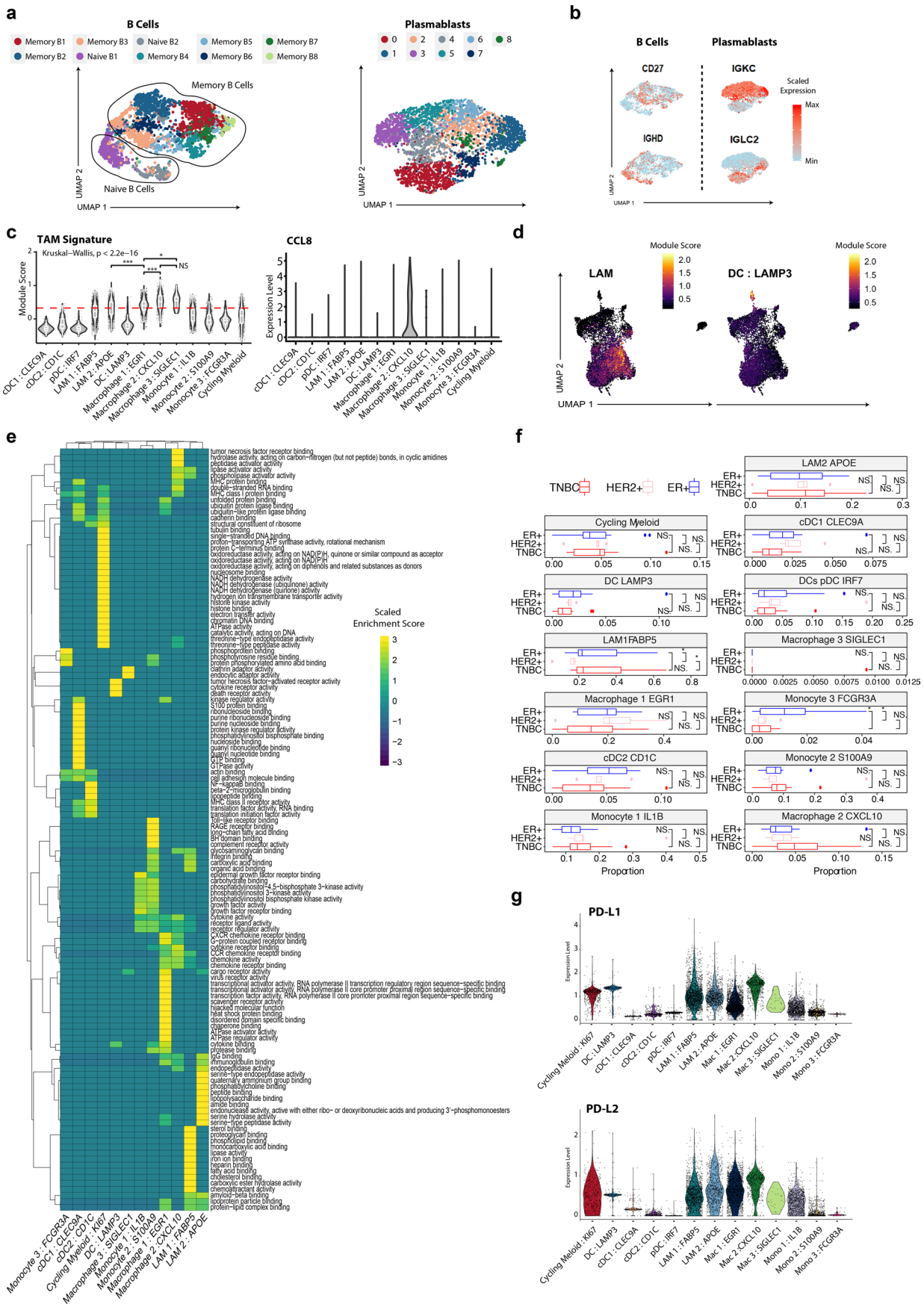
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Supplementary data for T-cells and innate lymphoid cells. a, Dotplot visualizing averaged expression of canonical markers across T-cell and innate lymphoid clusters. **b**, Cytotoxic and dysfunctional gene signature scores across T-cell and innate lymphoid clusters. A Kruskal-Wallis test was performed to compare significance between (pairwise two-sided t-test for each cluster compared to the mean, p-values denoted by asterisks: *p < 0.05, **p < 0.01, ***p < 0.001 and ****p < 0.0001). Red line indicates the median expression. **c**, Dysfunctional gene signature scores of CD8 : LAG3 and CD8+ T : IFNG clusters across clinical subtypes (n = 26; 11 TNBC, 10 ER+ and 5 HER2+). A pairwise two-sided t-test for each cluster was performed to determine significance. P-values denoted by asterisks: *p < 0.05, **p < 0.01, ***p < 0.001 and ****p < 0.0001. **d**, Differentially expressed immune modulator genes, stratified by T-cell and Myeloid clusters, compared across breast cancer subtypes. A pairwise MAST comparison was performed to obtain bonferroni corrected p-values. All genes displayed are statistically significant (p-value < 0.05). **e**, Pairwise two-sided t-test comparison of *LAG3*, *CD27*, PD-1 (*PDCD1*) and *CD70* log-normalised expression values in *LAG3*/c8 T-cells across breast cancer subtypes (n = 26; 11 TNBC, 10 ER+ and 5 HER2+). **f**, Enrichment of *PDCD1*, *CD27*, *LAG3* and *CD70* expression in the METABRIC cohort between clinical subtypes (n = 1,608; 209 Basal, 224 Her2, 700 LumA and 475 LumB). A pair-wise Wilcox test was performed to identify statistical significance. P-values denoted by asterisks: *p < 0.05, **p < 0.01, ***p < 0.001 and ****p < 0.0001. Box plots in **b** and **f** depict the first and third quartiles as the lower and upper bounds, respectively. The whiskers represent 1.5x the interquartile range and the centre depicts the median.



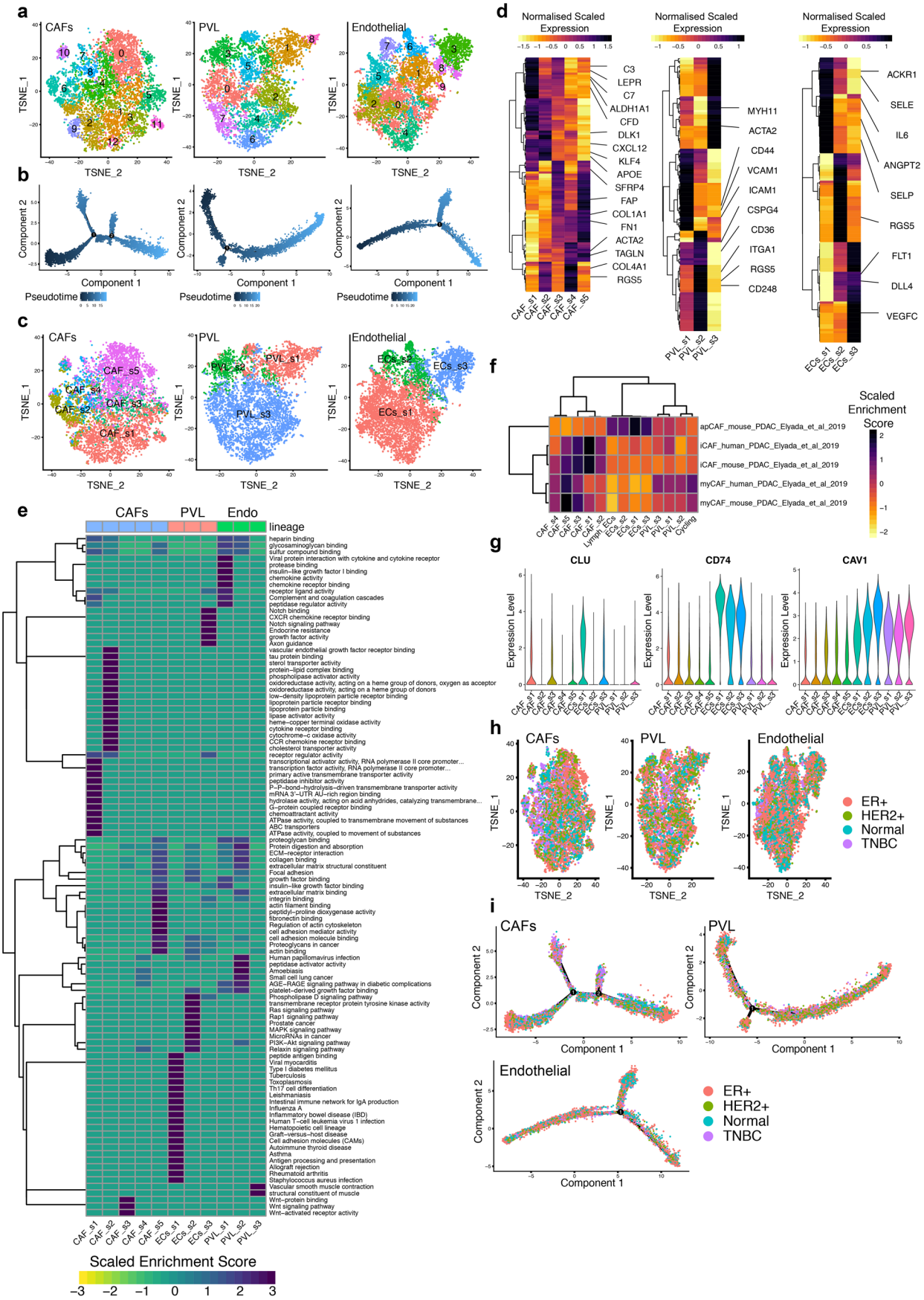
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Gene expression of immune cell surface receptors across malignant, immune and mesenchymal clusters and breast cancer clinical subtypes. a, Averaged expression and clustering of 133 clinically targetable receptor or ligand immune modulator markers across all cell types grouped by clinical breast cancer subtypes (TNBC, $HER2^+$ and ER^+). Gene lists were manually curated through systematic literature search of known immune modulating proteins expressed on the surface of cells. Default parameters for hierarchical clustering were used via the 'pheatmap' package for the visualization of gene expression values.



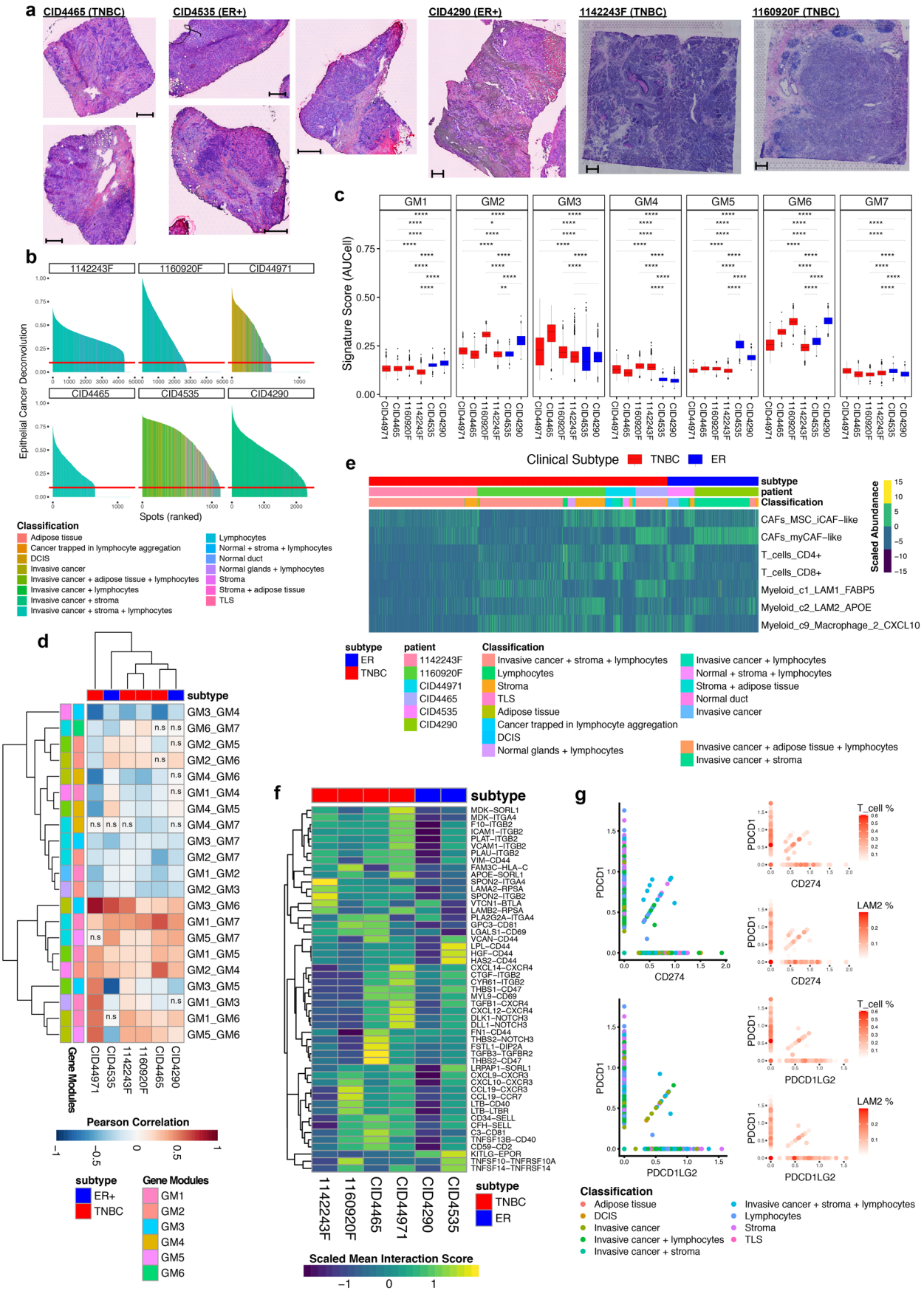
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Supplementary data for B-cells, Plasmablasts and Myeloid cells. a, UMAP visualization of all reclustered B-cells (n = 3,202 cells) and Plasmablasts (n = 3,525 cells) as annotated using canonical gene expression markers. **b**, Featureplots of *CD27, IGHD, IGKC* and *IGLC2* across naïve B cells, memory B cells, and Plasmablasts. **c**, Tumour associated macrophage (TAM) signature score obtained from Cassetta et al. 2019 and the expression of log-normalised levels of *CCL8* across all myeloid clusters (9,675 cells from 26 tumors). A pairwise two-sided t-test was performed to determine statistical significance for clusters of interest. P-values denoted by asterisks: *p < 0.05, **p < 0.01, ***p < 0.001 and ****p < 0.0001. Dashed red line marks median TAM module score or gene expression. A Kruskal-Wallis test was performed to compare significance between groups'. **d**, LAM and DC : LAMP3 gene expression signatures acquired from Jaitin et al. 2019 and Zhang et al. 2019 respectively, visualized on the myeloid UMAP clusters. **e**, Heatmap visualizing GO enrichment pathways across myeloid clusters. **f**, Proportion of myeloid clusters across clinical subtypes. Statistical significance was determined using a two-sided *t*-test in a pairwise comparison of means between groups (n = 26; 11 TNBC, 10 ER+ and 5 HER2+). P-values denoted by asterisks: *p < 0.05, **p < 0.01, ***p < 0.001 and ****p < 0.0001. **g**, Violin plots of imputed CITE-seq PD-L1 and PD-L2 expression values found on myeloid cells. Box plots in **c** and **f** depict the first and third quartiles as the lower and upper bounds, respectively. The whiskers represent 1.5x the interquartile range and the centre depicts the median.



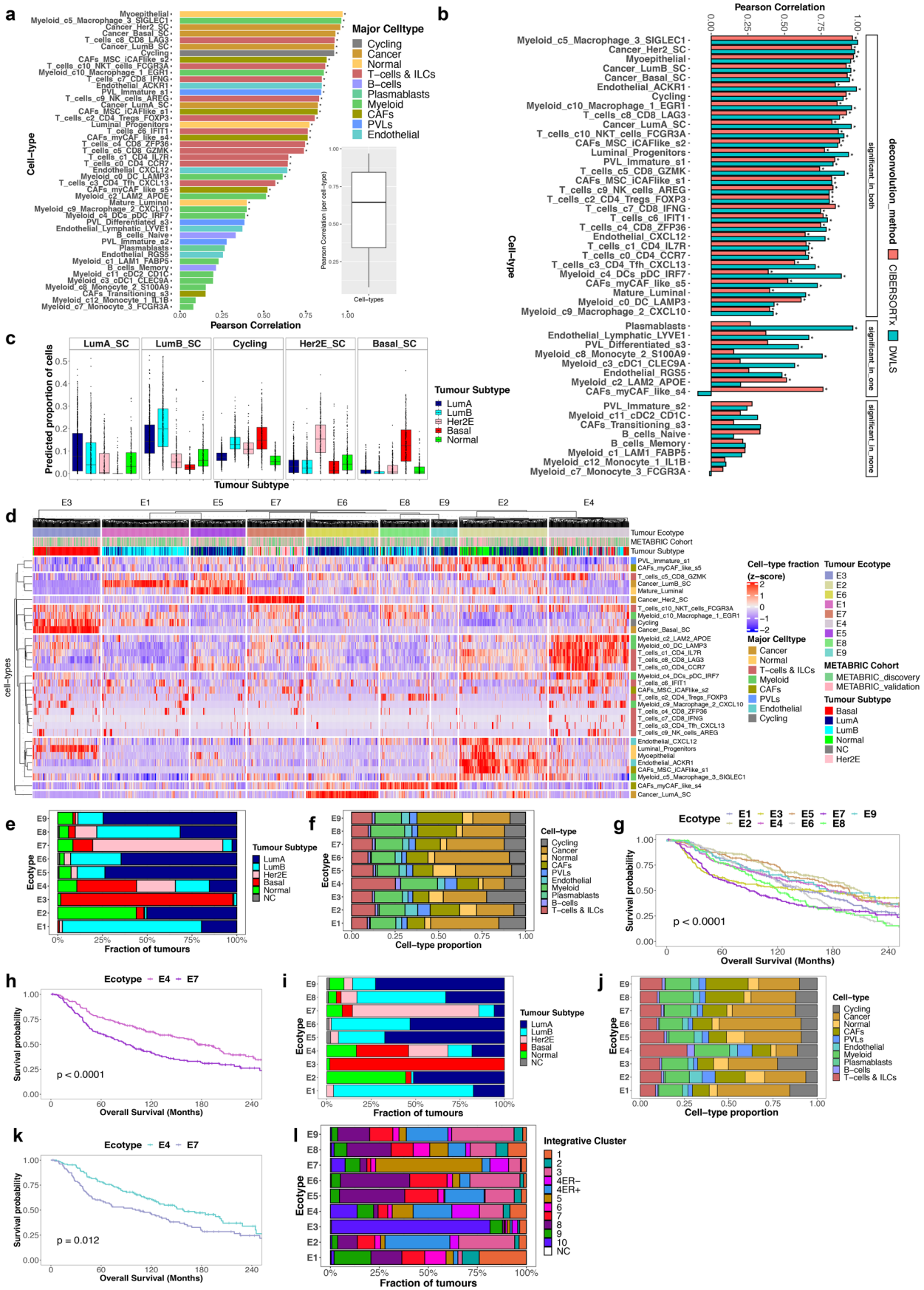
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Supplementary data for mesenchymal cell states and subclusters. a, *t*-SNEvisualization CAFs, PVL cells and endothelial cells using Seurat reclustered with default resolution parameters (0.8). **b**, Pseudotime plot for CAFs, PVL cells and endothelial cells, as determined using monocle. Coordinates are as in main Figs. 5c, 5e and 5g. **c**, *t*-SNE visualizations for CAFs, PVL cells and endothelial cells with monocle derived cell states overlaid. **d**, Heatmaps for CAFs, PVL cells and endothelial cells show cell state averaged log normalised expression values for all differentially expressed genes determined using the MAST method, with select stromal markers highlighted. **e**, Top 10 gene ontologies (GO) of each mesenchymal cell state, as determined using pathway enrichment with ClusterProfiler with all differentially expressed genes as input. **f**, Stromal cell state averaged signature scores for pancreatic ductal adenocarcinoma myofibroblast-like, inflammatory-like and antigen-presenting CAF sub-populations, as determined using AUCell. **g**, Enrichment of antigen-presenting CAF markers *CLU*, *CD74* and *CAV1* in various stromal cell states. **h**, Subclusters of CAFs, PVL cells and endothelial cells determined using Seurat show a strong integration with three normal breast tissue datasets, highlighting similarities in subclusters across disease status and clinical subtypes of breast cancer. **i**, Cell states of CAFs, PVL cells and endothelial cells determined using monocle show a strong integration with three normal breast tissue datasets and breast cancer clinical subtypes.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Supplementary data for spatial transcriptomics. a, H&E images for the remaining five breast tumors analysed using Visium (TNBC: CID4465, 1142243F and 1160920F; ER+: CID4535 and CID4290). Scale bars represent 500 μm. **b**, Histograms of cancer deconvolution values, as estimated using Stereoscope. Red line indicates the 10% cutoff used to select spots for scoring breast cancer gene-modules. Spots are colored by the pathology annotation. **c**, Box plots of gene module scores for all cancer filtered spots, as determined using AUCell, grouped by sample (TNBC=red; ER=blue). Statistical significance was determined using a two-sided *t*-test, with p-values adjusted using the Benjamini–Hochberg procedure. Box plots depict the first and third quartiles as the lower and upper bounds, respectively. The whiskers represent 1.5x the interquartile range and the centre depicts the median. P-values denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$ and ****$p < 0.0001$. **d**, Clustered gene module correlations across all cancer filtered spots. Color scales represent Pearson correlation values and are scaled per GM ('n.s' denotes not significant; two-sided correlation coefficient, Benjamini–Hochberg adjusted p-value < 0.05). **e**, Heatmap of the deconvolution values for inflammatory-like CAFs, myofibroblast-like CAFs, Macrophage *CXCL10*/c9, LAM1 and LAM2 clusters. Spots (columns) are grouped by sample and pathology. Deconvolution abundances (rows) are scaled by cell type. **f**, Predicted signaling in tissue spots enriched for iCAFs and CD4/CD8+ T-cells. Spots filtered for CAF-ligands and T-cell receptors detected by scRNA-Seq. The mean interaction scores of cell-signaling pairs are defined as the product of the ligand and receptor expression. **g**, Plots of PD-1 (*PDCD1*; y axis) expression with PD-L1 (*CD274*; x axis) or PD-L2 (*PDCD1LG2*; x axis) expression in spots enriched for CD4/CD8+ T-cells and LAM2 cells, as determined by Stereoscope. Abundance of CD4/CD8 T-cells (combined as T_cell here) and LAM2 are overlaid on the expression plots.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Supplementary figure for CIBERSORTx cell-type deconvolution. a, Bar and boxplot (inset) of the Pearson correlation for 45 cell-types between the actual cell-fractions captured by scRNA-Seq and the CIBERSORTx predicted fractions from pseudo-bulk expression profiles (*denotes significance $p < 0.05$, two-sided correlation coefficient). Inset box plot depicts the first and third quartiles as the lower and upper bounds, respectively. The whiskers represent 1.5x the interquartile range and the centre depicts the median. **b**, Barplot comparing the Pearson correlation for cell-types between the actual cell-fractions captured by scRNA-Seq and the CIBERSORTx (red) and DWLS (blue) predicted fractions from pseudo-bulk expression profiles (*denotes significance $p < 0.05$, two-sided correlation coefficient). **c**, Boxplot comparing the CIBERSORTx predicted SCSubtype and Cycling cell-fractions in each METABRIC tumor, stratified by PAM50 subtypes ($n = 1,608$; 209 Basal, 224 Her2, 700 LumA and 475 LumB). Box plots depicted as described in b. **d**, Heatmap of ecotypes formed from the common METABRIC tumors (columns) identified from combining ecotypes generated using CIBERSORTx with all 32 significantly correlated cell-types (rows), when using CIBERSORTx on pseudo-bulk samples. **e-f**, Relative proportion of the PAM50 subtypes (**e**) and major cell-types (**f**) in each ecotype, when combining CIBERSORTx consensus clustering results. **g-h**, Kaplan-Meier (KM) plot of all patients with common tumors in each of the ecotypes (**g**) and patients with tumors in ecotypes E4 and E7 (**h**), when combining CIBERSORTx consensus clustering results. p-values calculated using the log-rank test. **i-j**, Relative proportion of the PAM50 molecular subtypes (**i**) and major cell-types (**j**) of the common tumors from combining CIBERSORT and DWLS generated ecotypes. **k**, KM plot of the patients with tumors in ecotypes E4 and E7, formed from combining CIBERSORT and DWLS generated ecotypes. p-value calculated using the log-rank test. **l**, Relative proportion of the METABRIC integrative cluster annotations of the tumors in each ecotype, as determined using CIBERSORTx across all cell-types.

nature research

Corresponding author(s): A/Prof. Alexander Swarbrick

Last updated by author(s): Sep 11, 2020

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No specific software was used for the acquisition of the data.
Data analysis	Data analysis was performed using R (v3.5.0 and v3.6.1) with the following packages: Seurat (v3.0.0- for scRNA-Seq and v3.2.0 for ST analysis), DropletUtils (v1.2.2; for EmptyDrops), Garnett (v0.1.4), cluster (v2.0.7-1), CITE-seq-count (v.1.4.3), cola (v1.2.0), MAST (v1.12.0 - as part of DWLS method, v1.8.2 for differential gene expression), clusterProfiler (3.14.0), pheatmap (v1.0.12), Monocle (v2.10.1), STUtility (v0.1.0), AUCell (v1.4.1), survival (v2.44-1.1), survminer (v0.4.7), DWLS (code obtained from https://github.com/dtsoucas/DWLS on 30/11/2020). Additional software included Cellranger (v2.0, 10X Genomics), Space Ranger (v1.0.0, 10X Genomics), Loupe (v4.0.0, 10X Genomics) CIBERSORTx (v1.0), snakemake (v5.5.4), stereoscope (v0.2.0) and QuPath (v0.2.0). Description of the data analysis described in methods can be found at https://github.com/Swarbricklab-code/BrCa_cell_atlas.

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All processed scRNA-seq data is available for in-browser exploration and download through the Broad Single-Cell portal at https://singlecell.broadinstitute.org/single_cell/study/SCP1039. Raw scRNA-Seq data from this study has been deposited in the European Genome-Phenome Archive (EGA), which is hosted by the EBI and the CRG, under the accession code EGAS00001005173. All ST data from this study is available from the Zenodo data repository (DOI: 10.5281/zenodo.4739739).

Field-specific reporting

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

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

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

Life sciences study design

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

Sample size	Given the discovery and exploratory nature of our study, no statistical methods were applied to determine sample size. The number of tumor tissues analyzed in this study (26x scRNA-Seq and 6x spatial transcriptomics) reflect the availability and accessibility of the patient cohort to represent the major subtypes of breast cancer, as well as the funding limitations of the project grant.
Data exclusions	For filtering our scRNA-Seq datasets, we excluded cell barcodes determined as 'emptyDroplets' using the DropletUtils (v1.2.2) package. In addition, we excluded poor quality cells with genes and unique molecular identifier (UMI) counts less than 200 and 250, respectively, and a mitochondrial percentage less than 20%. For spatial data, all spatial barcodes that did not fall under tissue, as determined using the Space Ranger software (10X Genomics) were excluded. These steps are described in the manuscript
Replication	The single-cell RNA-Sequencing and spatial transcriptomics experiments described in this study consisted of a an independent single replicate per patient tumor. This was primarily due to the limited tissue samples collected from clinical specimens, as well as funding limitations, and is typical for this field
Randomization	Considering the exploratory nature of our clinical multi-omics study, randomization was not generally relevant for our study.
Blinding	The investigators had no prior knowledge of patient identities and clinical information prior to the collection of scRNA-Seq data. For scRNA-Seq data analysis, investigators were not blinded to this clinical information, as this was required to guide and design the analysis. For the spatial transcriptomics analysis, investigators were not blinded to clinical information prior to collection of data, as this was required for the selection of samples.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Details of all commercial antibodies used in this study can be found in Supplementary Table 11. All antibodies were used at a dilution of 1:100, as recommended by the manufacturer.

Validation	All commercial antibodies (Supplementary Table 11) used for CITE-Seq in this study were acquired from Biolegend and validated by the vendor by flow cytometry.

Human research participants

Policy information about studies involving human research participants

Population characteristics	All patients involved in this study were females aged between 35-88 with breast cancer across the three major clinical subtypes (ER+, HER2+ and TNBC). Most of the patients in our study had not undergone treatment. Detailed clinical information can be found in Supplementary Table 1.
Recruitment	Patients who fit the clinical criteria and consented to the study were selected for inclusion for multi-omics analysis. Our study had no self-selection bias or other biases in the recruitment of patients.
Ethics oversight	Primary untreated breast cancers (Supplementary Table 1) were collected with written consent from all patients under the protocols x13-0133, x19-0496, x16-018 and x17-155 with approval from all relevant human research ethics committees (Sydney Local Health District Ethics Committee, Royal Prince Alfred Hospital zone, and the St Vincent's hospital Ethics Committee). Consent included the use of all de-identified patient data for publication. Participants were not compensated.

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