Published in final edited form as:
Nat Genet. 2021 September ; 53(9): 1334–1347. doi:10.1038/s41588-021-00911-1.

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

Sunny Z. Wu[1,2,†], Ghamdan Al-Eryani[1,2,†], Daniel Roden[1,2,†], Simon Junankar[1,2], Kate Harvey[1], Alma Andersson[3], Aatish Thennavan[4], Chenfei Wang[5], James 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], Laurence Gluch[9], Mun N. Hui[1,10], Jane Beith[10], Andrew Parker[11], Elizabeth Robbins[12], Davendra Segara[11], Caroline Cooper[13,14], Cindy Mak[15], Belinda Chan[15], Sanjay Warrier[15,16], Florent Ginhoux[17], Ewan Millar[18], Joseph E. Powell[6,19], Stephen R. Williams[7], X. Shirley Liu[5], Sandra O'Toole[1,12], Elgene Lim[1,2,11], Joakim Lundeberg[3], Charles M. Perou[4], Alexander Swarbrick[1,2,*]

[1)]The Kinghorn Cancer Centre and Cancer Research Theme, Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia

[(2)]St Vincent's Clinical School, Faculty of Medicine, UNSW Sydney, NSW 2052, Australia

[(3)]Science for Life Laboratory, Division of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden

[(4)]Department of Genetics, University of North Carolina, Chapel Hill, NC, 27599, USA

[(5)]Department of Data Sciences, Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Harvard T.H. Chan School of Public Health

[*]Corresponding author: a.swarbrick@garvan.org.au.
[†]These authors contributed equally

Author Contributions
AS conceived the project and directed the study with input from all authors. SZW, AS, DR, GA-E and SJ wrote the manuscript with input from all authors. EL, SW, MNH, BC, CC, CM, DS, ER, AP, JB, SOT, EM and LG contributed to experimental design, procured patient tumor tissue and assisted with interpreting data. SZW, GA-E & KH performed single cell captures and KH analysed all clinical information. SZW, KH and GA-E optimized and performed tumor dissociation experiments. GA-E optimised and performed antibody staining for CITE-Seq experiments. NB and GA-E performed the CITE-Seq data processing. CL-C and SZW performed scRNA-Seq experiments on the Chromium Controller. CL-C helped perform the next-generation sequencing of the scRNA-Seq libraries. SZW performed the pre-processing, data-integration and reclustering steps for the scRNA-Seq data. JT performed the analysis and benchmarking of inferCNV. AT and CMP led the development of scSubtype. DR interpreted and led the analyses for breast cancer gene module analyses. KH and TW performed H&E and IHC experiments. SOT independently assessed and scored all histology in this study. GA-E interpreted and performed analyses of immune cells with intellectual input from SJ. C-AD and FG provided intellectual input related to myeloid cluster annotation. SZW interpreted and performed all analyses of stromal cells. DK and CL-C performed Visium experiments with input from JEP. VG helped perform pre-processing of Visium datasets. SRW, NIW, CRU, JGC and ZWB performed Visium experiments and data processing from an independent laboratory. AA performed the Stereoscope deconvolution with input from JL. SZW performed downstream analysis of Visium data with guidance from AA, LL, GA-E and JL. DR interpreted and performed CIBERSORTx analysis. SZW and DR performed survival analyses. CW and XSL provided intellectual input and guidance on bulk deconvolution and survival analyses.

Code Availability
Code related to the analysis in this study can be found on github at https://github.com/Swarbricklab-code/BrCa_cell_atlas[71].

Competing Interests
CMP is an equity stock holder and consultant of BioClassifier LLC; CMP is also listed as an inventor on patent applications for the Breast PAM50 Subtyping assay.
JL is an author on patents owned by Spatial Transcriptomics AB covering technology presented in this paper. The remaining authors declare no competing interests.

[6]Garvan-Weizmann Centre for Cellular Genomics, Garvan Institute of Medical Research, Sydney, Australia

[7]10x Genomics, Pleasanton, CA, 94588, USA

[8]Gustave Roussy Cancer Campus and Institut National de la Santé Et de la Recherche Médicale (INSERM) U1015, Equipe Labellisée—Ligue Nationale contre le Cancer, Villejuif, France

[9]The Strathfield Breast Centre, Strathfield, NSW 2135, Australia

[10]Chris O'Brien Lifehouse, Camperdown, NSW 2050, Australia

[11]St Vincent's Hospital, Darlinghurst, NSW 2010, Australia

[12]Royal Prince Alfred Hospital, Camperdown, NSW 2050, Australia

[13]Pathology Queensland, Princess Alexandra Hospital, Brisbane, Queensland 4102, Australia

[14]Southside Clinical Unit, Faculty of Medicine, University of Queensland, Brisbane, Queensland 4102, Australia

[15]Department of Breast Surgery, Chris O'Brien Lifehouse, NSW 2050, Australia

[16]Royal Prince Alfred Institute of Academic Surgery, Sydney University

[17]Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), BIOPOLIS, Singapore, Singapore

[18]NSW Health Pathology, Department of Anatomical Pathology, St George Hospital, Kogarah NSW, Australia, School of Medical Sciences, UNSW Sydney, Kensington, Australia, Faculty of Medicine & Health Sciences, Sydney Western University, Campbelltown NSW, Australia

[19]UNSW Cellular Genomics Futures Institute, University of New South Wales, Sydney, Australia

Abstract

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 remains limited. Here we present a single cell and spatially resolved transcriptomics analysis of human breast cancers. We develop a single cell method of intrinsic subtype classification (scSubtype) to reveal recurrent neoplastic cell heterogeneity. Immunophenotyping using CITE-Seq provides high-resolution immune profiles, including novel PD-L1/PD-L2+ macrophage populations associated with clinical outcome. Mesenchymal cells displayed diverse functions and cell surface protein expression through differentiation within 3 major lineages. Stromal-immune niches were spatially organized in tumors, offering insights into anti-tumor 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.

Keywords

breast cancer; single cell; spatial; genomics; transcriptomics; tumor heterogeneity; fibroblast; CAF; stroma; cancer immunology; immune

Introduction

Breast cancers are clinically stratified based on the expression of the estrogen receptor (ER), progesterone receptor (PR) and overexpression of 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 (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 (BLBC) and normal-like. There is ~70-80% concordance between molecular subtypes 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.

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, tumour associated macrophages (TAM) are often associated with poor prognosis[10] and are recognised 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 anti-tumor immunity[14,15]. However, progress has been impeded by a lack of a clear cellular taxonomy (recently reviewed in Sahai *et al.*[16]). Recent studies of cancer-associated fibroblasts (CAFs) identify 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 TME, and with disease status and progression remains 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 novel 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 analyse 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, there exists little data systematically mapping the spatial transcriptomic architecture of breast tumors, which can determine how cells in the TME are organized as functional units.

Results

A high-resolution cellular landscape of human breast cancers

To elucidate the cellular architecture of breast cancers, we analyzed 26 primary pre-treatment 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], 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. 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; 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 single cell RNA-Seq data

As 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 “pseudo-bulk” 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 pseudo-bulk samples with the TCGA dataset of 1,100 breast tumors using an ~2,000 gene intrinsic breast cancer genelist[3] (Extended Data Fig. 2a–b). Training samples were selected from those with concordance between pseudo-bulk 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; 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 pseudo-bulk calls in the 10 training set samples and 66% agreement on the test set samples (Extended Data Fig. 2d; Supplementary Table 3). Of the 3 test set disagreements, two were LumA vs LumB, which are related profiles that may be 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 testing 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 the bulk tumor RNA-Seq profile 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 pseudo-bulk profiles (Extended Data Fig. 2a–c). These results highlight the strong concordance between our three subtyping methods when applied across bulk and scRNA-Seq datasets.

scSubtype revealed that 13/20 samples had less than 90% of neoplastic cells falling under one molecular subtype, while only one 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 IHC in two ER+ cases which 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 is 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; Extended Data Fig. 2d), which is consistent with the clinical IHC result. Bulk and pseudo-bulk 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; 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 ‘bulk’ molecular subtype to develop a classifier. To complement this, we investigated the biological pathways driving intra-tumor transcriptional heterogeneity (ITTH) in an unsupervised manner, using integrative clustering of tumours with at least 50 neoplastic cells, to generate 574 gene-signatures of ITTH. These gene-signatures identified 7 robust groups, “gene-modules”, based on their Jaccard similarity (Extended Data Fig. 3a). Each gene-module (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 (e.g., E2F_TARGETS), driven by genes including *MKI67, PCNA* and *CDK1*. GM3 was predominately 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, TNFa and 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 7 GMs and used hierarchical clustering to identify cellular correlations (Extended Data Fig. 3c). This clearly separated neoplastic cells into groups, reducing the large inter-tumor 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 significantly associated with scSubtype calls, whereas others displayed more diverse subtype associations (Fig. 2f–g; Extended Data Fig. 3e–f). Cells assigned to GM1 and GM5 were predominantly enriched for the luminal subtype, where GM1 was almost exclusively composed of LumA cells and GM5 was mostly composed of LumB cells. As 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 gene module-based cell state assignments to get a view into the intra-tumour heterogeneity of the neoplastic cells. Similar to scSubtype (Fig. 2b), we saw evidence for cellular heterogeneity that broadly aligns with, but not constrained by, the subtype of the tumor (Fig. 2h). scSubtype and gene module analysis provide complementary new approaches to classifying neoplastic ITTH and 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 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 were comprised of *FOXP3*+ regulatory T cells (T-Regs) marked by CD25 protein expression (CD4+ T-cells:FOXP3/c2), T follicular helper (Tfh) cells (*CXCL13, IL21* and *PDCD1*; CD4+ T-cells:CXCL13/c3), naïve/central memory CD4+ (CD4+ T-cells:*CCR7*/c0), and a Th1 CD4 T effector memory (EM) cluster (CD4+ T-cells:*IL7R*/c1) (Fig. 3b; Extended Data Fig. 5a). Of the five CD8 clusters, three were comprised 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) signature (*SG15, IFIT1* and *OAS1*; T-cells:*IFIT1*/c6) and proliferation (T-cells:*MKI67*/c11) were identified, both comprised of CD4+ and CD8+ T-cells. We also identified NK cells (NK cells:*AREG*/c9) and NKT-like cells (NKT cells:*FCGR3A*/c10) by their expression of αβ T-cell receptor and NK markers (*KLRC1, KLRB1, NKG7)* (Fig. 3b; Extended Data Fig. 5a).

Consistent with the enrichment of TILs and CD8+ T-cells in TNBC[37], T cell clusters *IFIT1*/c6, *LAG3*/c8 and *MKI67*/c11 made up a higher proportion in 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; Extended Data Fig. 5b–c). Furthermore, luminal and HER2+ tumors tended to have checkpoint molecule expression distinct from TNBC (Fig. 4f; 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; Extended Data Fig. 5e). We examined the expression of *PDCD1, CD27* and *CD70* in the METABRIC[40] and TCGA[32] cohorts, 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 BCR specific gene segments rather than variable biological gene expression programs.

Myeloid Cells

Myeloid cells formed 13 clusters which could be identified in all tumors at varying frequencies (Fig. 4a). No granulocytes were detected, likely due to their sensitivity to tumour dissociation protocols and their low abundance[22,41,42]. Monocytes formed 3 clusters: Mono:*IL1B*/c12; Mono:*S100A9*/c8; and Mono:*FCGR3A*/c7, with the Mono:*FCGR3A* population forming a small distinct cluster characterized by high CD16 protein expression (Fig. 4b–c). We identified conventional dendritic cells (cDC) that expressed either *CLEC9A* (cDC1:*CLEC9A*/c3) or *CD1C* (cDC2:*CD1C*/c11); plasmacytoid DC (pDC) 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). Macrophages formed 6 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. 6c)[10]. Notably, we identified two novel macrophage populations (LAM1:*FABP5/*c1 and LAM2:*APOE/*c2) outside of the conventional “M1/M2” classification that comprised 30-40% of the total myeloid cells (Fig. 4a–c). These cells bear close transcriptomic similarity to a recently described population of lipid-associated macrophages (LAM) 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; 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 LAM 1:*FABP5* cells in the HER2+ tumors (Fig. 4d; 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 LAM 1:*FABP5* signature correlates with worse survival (Fig. 4e). While the RNA encoding PD-L1 (*CD274*) and PD-L2 (*PDCD1LG2*) were highly co-expressed by the Mac:*CXCL10* and DC:*LAMP3* myeloid populations (Fig. 4f), analysis of 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* (Fig. 4b; 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; Extended Data Fig. 8a) including CAFs (*PDGFRA* and *COL1A1;* Fig. 5c–d), perivascular-like cells (PVL; *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[46] revealed five CAF states (Fig. 5c; Extended Data Fig. 8b–c). State 1 (referred to as s1 herein) had features of mesenchymal stem cells (MSC) and inflammatory-like fibroblasts (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 towards differentiated states s4 and s5, which resembled myofibroblast-like (myCAF) states through the increased expression of *ACTA2* (αSMA), *TAGLN, FAP* and *COL1A1*[15] and the enrichment of ECM-related pathways. Previously reported iCAF and myCAF 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 (apCAFs) signatures, however, selected apCAF markers *CD74, CLU* and *CAV1* were broadly expressed across all stromal cells (Extended Data Fig. 8g).

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 PDGF 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 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 myCAF gene signatures (Extended Data Fig. 8f), and their shared expression of the CAF marker *ACTA2* suggest that PVL s3 cells have been 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/MHC (*HLA-DRA*) (Extended Data Fig. 8d–e). These markers decreased along pseudotime as cells branched into two states, which both had elevated expression of *DLL4*, a marker reported for endothelial tip-like cells (Fig. 5h)[49,50]. Endothelial s2 was distinguished by *RGS5* and *ESM1*, whilst s3 expressed regulators of cell migration and angiogenesis (*CXCL12* and *VEGFC*)[51]. As 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 markers *ACRK1, 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 likely 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 (ST) on six samples ("local cohort") comprising two ER+ (CID4535 and CID4290) and two TNBC (CID44971 and CID4465) from our scRNA-Seq cohort, and two additional TNBC (1142243F and 1160920F) processed in an independent laboratory (Fig. 6a; Extended Data Fig. 9a). To deconvolute the cellular composition of each ~55 uM diameter spot, we applied a probabilistic model called Stereoscope[54] using clinical subtype matched scRNA-Seq data. Cell types were found associated with their appropriate pathological annotation (Fig. 6b).

We earlier showed that gene modules were enriched for distinct microenvironment associated pathways and factors, and thus, we hypothesised that gene modules would be spatially organised 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 gene module 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; Extended Data Fig. 9c). These data suggest that these gene modules are not an artefact of dissociation-based methodology. To systematically understand the spatial relationship between modules, we computed Pearson correlations between gene module scores in all cancer locations. This revealed two major clusters that mostly conserved across all six cases, including GM1, GM3, GM5 and GM6 in one cluster, and GM2 and GM4 in the other (Fig. 6d; 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 novel heterotypic cellular interactions

While several studies have shown an important role for mesenchymal cells in regulating anti-tumor immunity[14,55], interactions between stromal and immune cells have yet to be profiled in tissues. Deconvolution revealed spatially distinct subclasses of CAFs, with myCAFs (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 myCAFs 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[54], suggesting that this relationship is conserved across clinical subtypes (Fig. 7a). Consistent with the immunoregulatory properties of iCAFs described above, iCAFs co-localized with several lymphocyte populations across both studies, including memory/naive B-cells and CD4+/CD8+ T-cells (Fig. 7a; Fig. 7d–e). MyCAFs 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 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), complement pathway, transforming growth factor beta (TGFB1/TGFB3-TGFBR2) and lymphocyte inhibitory/activation molecules (LTB-LTBR, TNFSF14-LTBR and LTB-CD40, VTCN1/B7H4-BTLA) (Fig. 7f; 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 were also found in areas with high stroma, adipose and lymphocytes by morphology (Extended Data Fig. 9e). LAM1 and LAM2 cells show a modest negative spatial correlation with each other in most cases, which might indicate that a common LAM cell is polarised towards 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 8 tumors across all three subtypes (Fig. 7a). Spots enriched for LAM2 cells and CD4+/CD8+ T-cells across multiple tumors co-expressed PD-L1-PD-1 (*CD274-PDCD1*) and PD-L2-PD-1 (*PDCD1LG2-PDCD1*), suggesting these cells likely have functional relevance in immunoregulation (Extended Data Fig. 9g). In addition, positive Pearson correlations were identified between Mac:CXCL10/c9 cells and CD8 T-cells across a majority of cases (Fig. 7a), which were mostly enriched in spots annotated as ‘Invasive cancer + stroma + lymphocytes’ (Fig. 7g; Extended Data Fig. 9e), suggesting these niches may have functional relevance in regulating anti-tumor immunity.

Breast tumor ecotypes associated with patient survival

Our single cell data has 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 by using our single-cell signatures with CIBERSORTx[56]. Estimating cell fractions from pseudo-bulk samples generated from our single-cell datasets showed good overall correlation between the captured cell-fractions and the predicted proportions (median correlation ~0.64), with a majority (32) of cell-types showing a significant correlation (Extended Data Fig. 10a). An alternative deconvolution method, DWLS[57], 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 (Wilcox test, $p<2.2e-16$) of the four scSubtypes in tumors with matching bulk-PAM50 classifications. Significant enrichment (Wilcox test, $p<2.2e-16$) of cycling cells in Basal, LumB and HER2E tumors was also shown (Extended Data Fig. 10c). Consensus clustering revealed 9 tumor clusters with similar estimated cellular composition ("Ecotypes") (Fig. 8a). 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 anti-tumor immunity (Fig. 8a), including exhausted CD8 T cells (*LAG3*/c8), along with Th1− (*IL7R*/c1) and central memory (*CCR7*/c0) CD4 T cells. 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 associated with poor 5-year survival (Fig. 8d), 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 anti-tumor immune cells.

To further assess ecotype robustness, we repeated the consensus clustering using only the 32 significantly correlated cell-types, as well as the DWLS method. Substantial overlap of tumours (Supplementary Table 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 (int-clusters) identified by METABRIC[40] (Extended Data Fig. 10l). E3 had a high proportion of cancers from int-cluster 10, which also predominantly consists of basal-like tumours with similarly poor 5-year survival. E7 had a high proportion of *ERBB2* amplified and Her2E int-cluster 5 tumours. These are the worst prognosis groups in both the METABRIC and ecotype analysis. However, a majority of ecotypes don't clearly associate with a specific int-cluster or PAM50 subtype, reflected by the role of the 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

We provide here important advances toward an integrated cellular model for breast cancer classification. We define the cellular architecture of breast tumors at 3 levels: first, a detailed cellular 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 observe groups of tumors with similar cell type proportions and prognostic associations, that we name 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 under-represented. We have addressed this in part by using spatial transcriptomics on intact tissues. Future work may apply complementary technologies such as single-nuclei 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 identify 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 upon local interactions. To classify tumor cells in a manner consistent with the prior PAM50 bulk classifier, we developed scSubtype, which was able to subtype tumors with low cellularity, for which bulk analysis had failed. Although heterogeneous expression of subtype markers (e.g. 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 a majority of cancers. As for all classification methods, the performance of scSubtype will improve upon 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 intra-tumoral heterogeneity for intrinsic subtype may predict innate resistance to therapy and early relapse following therapy. For instance, the presence of basal-like or HER2-like cells in clinically luminal cancers (Fig. 2c) may cause early relapse following endocrine therapy.

We also conducted an integrative analysis to discover the gene expression programs underlying ITTH. This revealed that module 3 (EMT, IFN, MHC) and module 4 (proliferation) were mutually exclusive, suggesting that a mesenchymal-like state and proliferation are incompatible at cellular resolution. Furthermore, analysis of spatial data reveals 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[58]. 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 lipid-associated macrophages[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 *Trem2*-high expressing myeloid cells have an immunosuppressive role in mouse models of cancer[59,60], with human IHC analyses showing TREM2 expression in multiple subsets of macrophages in TNBC and an association with worse prognosis[60]. Our data extends upon these works by providing high resolution scRNA-Seq, cell surface protein and spatial characterisation 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 show 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 reveals three major cell populations, endothelial, CAF and PVL cells, consisting of 3-5 identifiable states each. Previous studies have shown that CAF states are interconvertible upon distinct tumor culture conditions, suggesting that this differentiation may also occur bi-directionally depending on external factors[17,18]. While differentiation from other progenitors like mesenchymal stem cells is possible, our pseudo-temporal 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 to 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 use deconvolution to define nine ecotypes amongst 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 partially associated with intrinsic subtype[4] and genomic classifiers[40], they are not simply surrogates for previous methods stratification. 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 personalise therapy.

Methods

Patient material, ethics and consent for publication

Primary untreated breast cancers used in this study (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.

Tissue dissociation

Samples were analyzed from fresh surgical resections and cryopreserved tissue[61]. Tumors were dissociated using Human Tumor Dissociation Kit (Miltenyi Biotec) following the manufacturer’s protocol. Where viability was < 80%, viability enrichment was performed using the EasySep Dead Cell Removal (Annexin V) Kit (StemCell Technologies) as per manufacturer’s protocol.

Single-cell RNA Sequencing 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 pair-ended 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 v2.0 software (10X Genomics). The EmptyDrops method from the DropletUtils package (v1.2.2)[62] was applied for cell filtering with additional cutoffs for cells with a gene and unique molecular identifier (UMIs) count greater than 200 and 250, respectively, and a mitochondrial percentage less than 20%. We used the Seurat v3.0.0 method[36] in *R* (v3.5.0) for data normalisation, dimensionality reduction and clustering using default parameters. Cell clusters were annotated using the Garnett method[29] (v0.1.4) with a classifier derived breast epithelial cell signatures[28], and immune and stromal cell types from

XCell[27]. Data integration was performed using Seurat v3.0.0[36] (see Supplementary Note for specific parameters used).

Identifying neoplastic from normal breast epithelial cells

CNV signal for individual cells was estimated using the inferCNV method (v0.99.7) with a 100 gene sliding window. Genes with a mean count of less than 0.1 across all cells were filtered out prior to analysis, and signal was denoised using a dynamic threshold of 1.3 standard deviations 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 (non-neoplastic), neoplastic or unassigned using a similar method to that previously described by Neftel *et al.*[30]. Briefly, inferred changes at each genomic loci were scaled (between −1 and +1) and the mean of the squares of these values were 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 (PAM) clustering was performed using the ‘pamk’ function in the *R* package ‘cluster’ (v2.0.7-1) to choose the optimum value for k (between 2-4) using silhouette scores, and the ‘pam’ function to apply the clustering. Thresholds defining normal and neoplastic cells were set at 2 cluster standard deviations to the left and 1.5 standard deviations below the first cancer cluster means. For tumors where PAM could not define more than 1 cluster, the thresholds were set at 1 standard deviation to the left and 1.25 standard deviations below the cluster means. This method was used to identify 27,506 neoplastic and 6084 normal cells in all tumors, the remaining 3208 cells were classed as unassigned. Only tumours with at least 200 epithelial cells were used for this neoplastic cell classification step.

Calling PAM50 on pseudo-bulks and matching bulk RNA-Seq

For calling molecular subtypes using the PAM50 method[3], we processed “pseudo-bulk” expression profiles for each tumor, named “Allcells-Pseudobulk”, in a similar manner to any bulk RNA-Seq sample (i.e. upper quartile normalized-log transformed). Prior to 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.*[63]. 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 FFPE blocks using the High Pure RNA Paraffin Kit (Roche #03 270 289 001). Libraries were sequenced on the HiSeq 2500 platform (Illumina) with 50 bp paired end reads. Transcript quantification was performed using Salmon[64]. We then called PAM50 on each bulk tumor using Zhao *et al.*[63] 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. Here, we identified robust training set samples using two subtyping approaches: (i) PAM50 subtyping of the *Allcells-Pseudobulk* datasets (described above); and (ii) hierarchical clustering of the *Allcells-*

Pseudobulk data with the 1,100 tumors in the TCGA breast cancer RNA-Seq dataset[32] using ~2000 genes from an intrinsic breast cancer genelist[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 is a histological subtype not used in the original PAM50 training set. Only tumor cells with greater than 500 UMIs were used for training and test datasets in scSubtype (total of 24,889 cells). Within each training set subtype, 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 the Seurat v3.0.0 package[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 weights 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 define as "scSubtype" gene signatures (Supplementary Table 4). To assign a subtype call to a cell we calculated the average (i.e. mean) read counts for each of the 4 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 ~20 thousand genes[33]. Averaged normalised 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 hrs 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 4uM. Sections were stained with Haematoxylin and Eosin for standard histological analysis. Immunohistochemistry (IHC) was performed on serial sections with pre-diluted primary antibodies against ER (clone 6F11; leica PA0151) or CK5 (clone XM26; leica PA0468) using suggested protocols on the BOND RX Autostainer (Leica, Germany). Antigen retrieval was performed for 20 min using

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). Slides were imaged using the Aperio CS2 Digital Pathology Slide Scanner and processed using QuPath (v0.2.0).

Gene module analysis of neoplastic intra-tumor heterogeneity

For each individual tumor, with more than 50 neoplastic cells, the neoplastic cells were clustered using Seurat v3.0.0[36] at five resolutions (0.4, 0.8, 1.2, 1.6, 2.0). MAST[65] (v1.12.0) was then used to identify the top-200 differentially regulated genes in each cluster. Only gene-signatures containing greater than 5 genes and originating from clusters of more than 5 cells were kept. In addition, redundancy was reduced by comparing all pairs of signatures within each sample and removing the pair with fewest genes from those pairs with a Jaccard index greater than 0.75. Across all tumors, a total of 574 gene-signatures of intra-tumor heterogeneity were identified.

Consensus clustering (using spherical k-means, skmeans, implemented in the cola R package (v1.2.0): https://www.bioconductor.org/packages/release/bioc/html/cola.html) of the Jaccard similarities between these gene-signatures was used to identify 7 robust groups, or gene-modules. For each of these, a gene module was defined by taking the 200 genes that had the highest frequency of occurrence across clusters and individual tumors. These are defined as gene-modules GM1 to GM7. A gene-module signature was calculated for each cell using AUCell[66] and each neoplastic cell was assigned to a module, using the maximum of the scaled AUCell gene-module signature scores. This resulted in 4,368, 3,288, 2,951, 4,326, 3,931, 2,500, 3,125 cells assigned to GM1 to GM7, respectively. These are defined as gene-module based neoplastic cell states.

Differential expression, module and pathway enrichment

Differential gene expression was performed using the MAST method[65] (v1.8.2). All DEGs from each cluster (log fold change greater than 0.5, p-value threshold of 0.05, and adjusted p-value threshold of 0.05; Supplementary Table 9 and 10) were used as input into the ClusterProfiler package[67] (v3.14.0) for gene ontology functional enrichment. Results were clustered, scaled and visualised using the pheatmap package (v1.0.12). Cytotoxic, TAM and Dysfunctional T-cell gene expression signatures were assigned using the *AddModuleScore* function in Seurat v3.0.0[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 which 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, PVL and Endothelial cells) using the Monocle 2 method[46] (v2.10.1) with default parameters as recommended by developers. Integrated gene expression matrices from each cell type were first exported from Seurat v3 into Monocle to construct a CellDataSet. All variable genes defined by the differentialGeneTest function (q-val cutoff < 0.001) were used for cell ordering with the setOrderingFilter function. Dimensionality reduction was performed with no normalisation methods and the DDRTree reduction method in the reduceDimension step.

CITE-Seq antibody staining

Samples were stained with 10X Chromium 3' mRNA capture compatible TotalSeq-A antibodies (Biolegend, USA). 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 (CID383) and two TNBC (CID4515 and CID3956). Staining was performed as previously described by Stoeckius et. al[35]. Briefly, a maximum of 1 million cells per sample was resuspended in 120 ul of cell staining buffer (Biolegend, USA) with 5 ul of Fc receptor Block (TrueStain FcX, Bioelegend, USA) for 15 min. This was followed by a 30 min staining of the antibodies at 4°C. A concentration of 1 ug / 100 ul was used for all antibody markers used in this study. The cells were then washed 3 times with PBS containing 10% FCS media followed by centrifugation (300 x g for 5min at 4°C) and expungement of supernatant.

CITE-Seq data processing and imputation

Demultiplexed reads were assigned to individual cells and antibodies with python package CITE-seq-count v.1.4.3 (https://github.com/Hoohm/CITE-seq-Count/tree/1.4.2). CITE counts were normalised and scaled with Seurat v.3.1.4. Imputation of CITE data was performed per individual cell type (B-cells, T-cells, myeloid cells, 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 OCT and stored at −80°C. Tissue blocks were cut into 10 μm sections and processed using Visium Spatial Gene Expression Kit (10X Genomics) according to manufacturer's instructions. First, breast tissue permeabilization condition was optimised using Visium Spatial Tissue Optimisation kit, which was found to be ideal at 12 minutes. Sections were H&E stained and imaged using a Leica microscope DM6000 (Leica, DE) under a 20x lens magnification, then processed for ST. The resulting cDNA library was checked for quality control, then sequenced using on an Illumina NovaSeq 6000 (illumina, US). 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 (v4.0.0) software (10X Genomics).

Visium spatial transcriptomics data processing

Reads were demultiplexed and mapped to the reference genome GRCh38 using the Space Ranger Software v1.0.0 (10X Genomics). Count matrices were loaded into the Seurat v3.2.0 and STutility (v0.1.0) packages for all subsequent data filtering, normalisation, filtering, dimensional reduction and visualization. Data normalisation was performed on independent tissue sections using the variance stabilizing transformation method implemented in the

SCTransform function in Seurat. We applied non-negative matrix factorization (NMF) to the normalised expression matrix using STutility (nfactors = 20).

Spatial deconvolution using Stereoscope

We performed deconvolution of spatial tissue locations using the Stereoscope[54] (v0.2.0), a probabilistic model for estimating cell type proportions using annotated scRNA-Seq data as input. Stereoscope was performed using default parameters (see Supplementary Note for greater details). We matched spatial and single cell data with respect to breast cancer clinical subtype. We deconvolved cell types across three tiers of classification including the major, minor and subset lineages.

Mapping cancer heterogeneity and cell signalling predictions

For investigating breast cancer gene modules, we first filtered for all spots where cancer epithelial cells were called using the Stereoscope method with a filter of 10%. Gene module gene lists were then scored using the AUCell method[66] (v1.4.1). Gene module correlations were then computed using Pearson correlation across all spots using *R* (*cor.test* function; p-value cutoff of 0.05). For cell-cell colocalizations across all tissue domains, we included 7 additional HER2+ datasets generated on a platform similar to the Visium[68]. In total, Pearson correlation was computed from the cell abundances across the tissue locations from 13 patients using *R* (*cor.test* function; p-value cutoff of 0.05). For cell signalling 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-signalling sets[69,70], and only ligands and receptors differentially expressed by iCAFs and CD4/CD8+ T-cells in the scRNA-Seq data, respectively (MAST; avg.logfc threshold 0.1). All regions annotated as normal ductal by pathology were also excluded for the above analyses.

Survival analysis of scRNA-Seq signatures

To assess impact of particular cell types described by scRNA-Seq (e.g. 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%, and survival curves were generated using the Kaplan Meier method with the ‘survival’ package (v2.44-1.1). We assessed the significance between two groups using the log-rank test statistics.

Tumor ecotype analysis using deconvolution

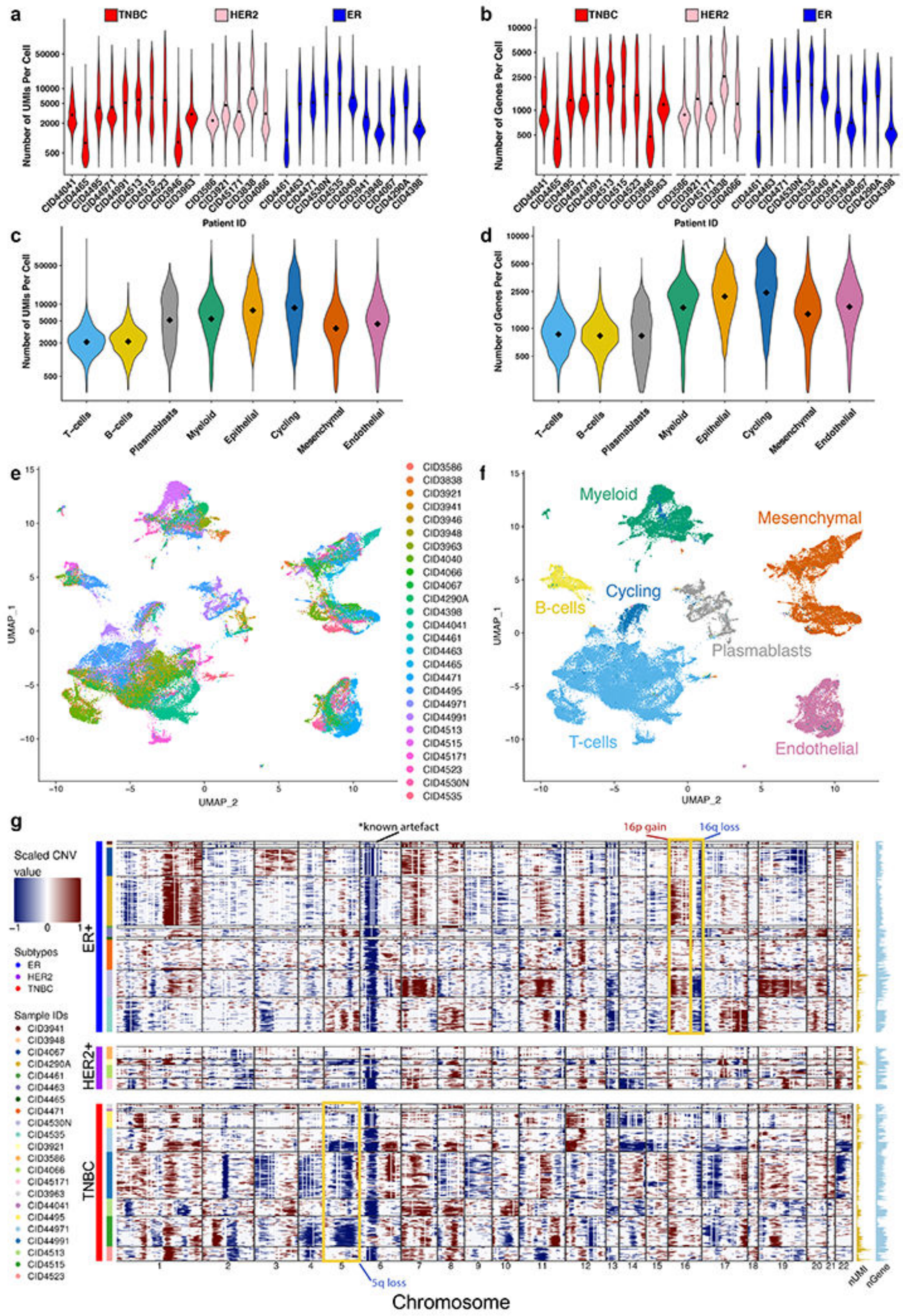
CIBERSORTx[56] (v1.0) and DWLS[57] were used to deconvolute predicted cell-fractions from a number of bulk transcript profiling datasets (see Supplementary Note for specific parameters). To prevent confounding of cycling cell-types we first assigned all neoplastic epithelial cells with a proliferation score > 0 as cycling and then combined these with “cycling” cell states from all other cell-types to generate a single

“Cycling” cell-state. Normalised METABRIC expression matrices, clinical information and PAM50 subtype classifications were obtained from https://www.cbioportal.org/study/summary?id=brca_metabric. Tumor ecotypes in the METABRIC cohort were identified using spherical k-means (skmeans) based consensus clustering (as implemented in the cola R package v1.2.0) of the predicted cell-fraction from either CIBERSORTx or DWLS, in each bulk METABRIC patient tumor. When comparing ecotypes between methods (i.e., consensus clustering results from using cell-abundances of all cell-types or just the 32 significantly significantly correlated cell-types from CIBERSORTx deconvolution and consensus clustering results from CIBERSORTx or DWLS cell-abundances) the number of tumour ecotypes was fixed as 9 and the tumour overlaps between all ecotype pairs was calculated (Supplementary Table 7 and 8). Common ecotypes were then identified by identifying the ecotype pairs with the largest average METABRIC tumour overlap. Differences in survival between ecotypes were assessed using Kaplan-Meier analysis and log-rank test statistics, using the *survival* (v2.44-1.1) and *survminer* (v0.4.7) R packages.

Statistics and Reproducibility

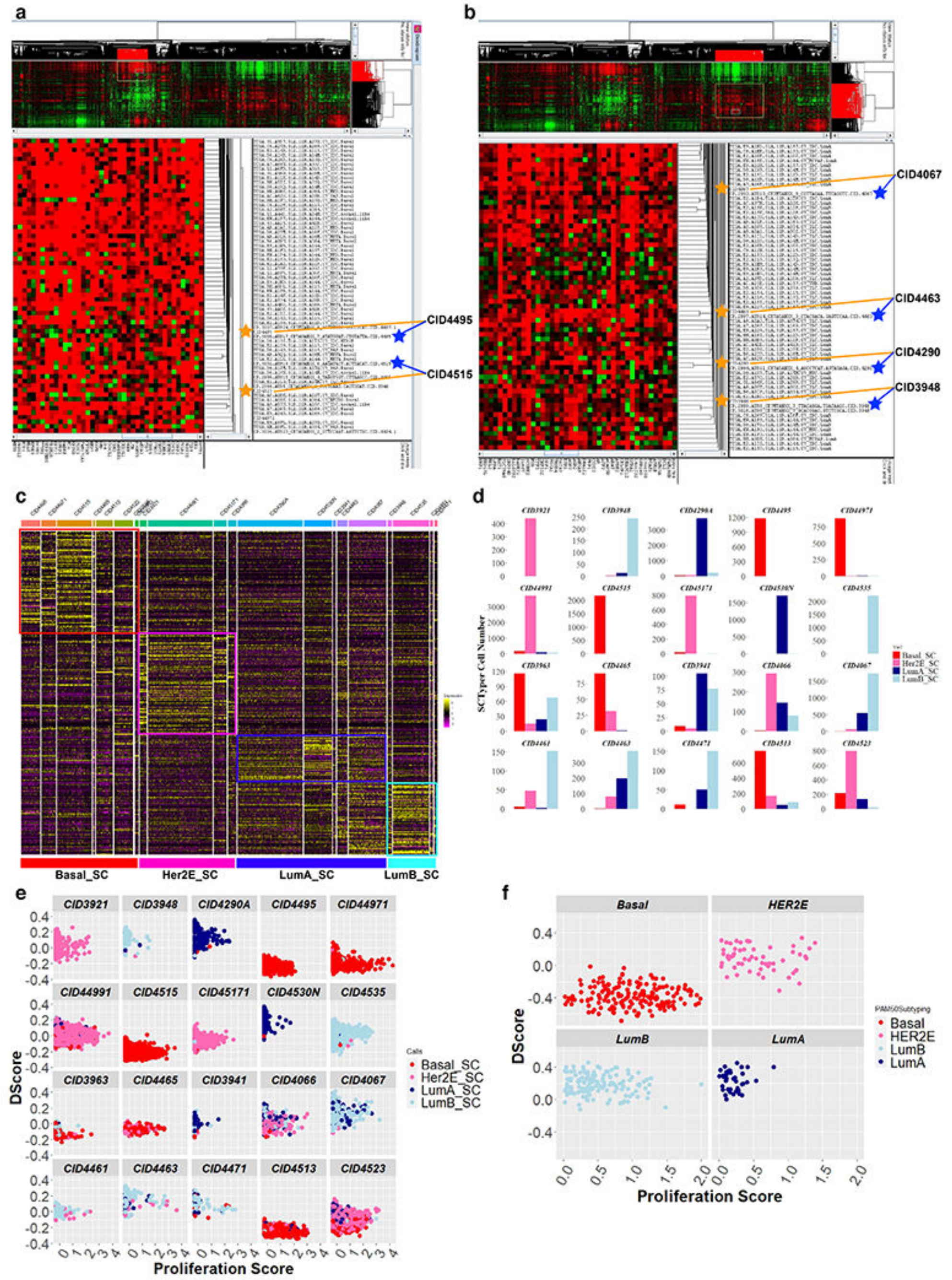
No statistical method was used to predetermine sample size. Statistical significance for differentially expressed genes were determined using the Wilcoxon Rank Sum test, with all p-values adjusted using bonferroni correction. All boxplots depict the first and third quartiles as the lower and upper bounds, respectively. The whiskers represent 1.5x the interquartile range and the centre depicts the median. All statistical tests used are defined in the figure legends.

Extended Data



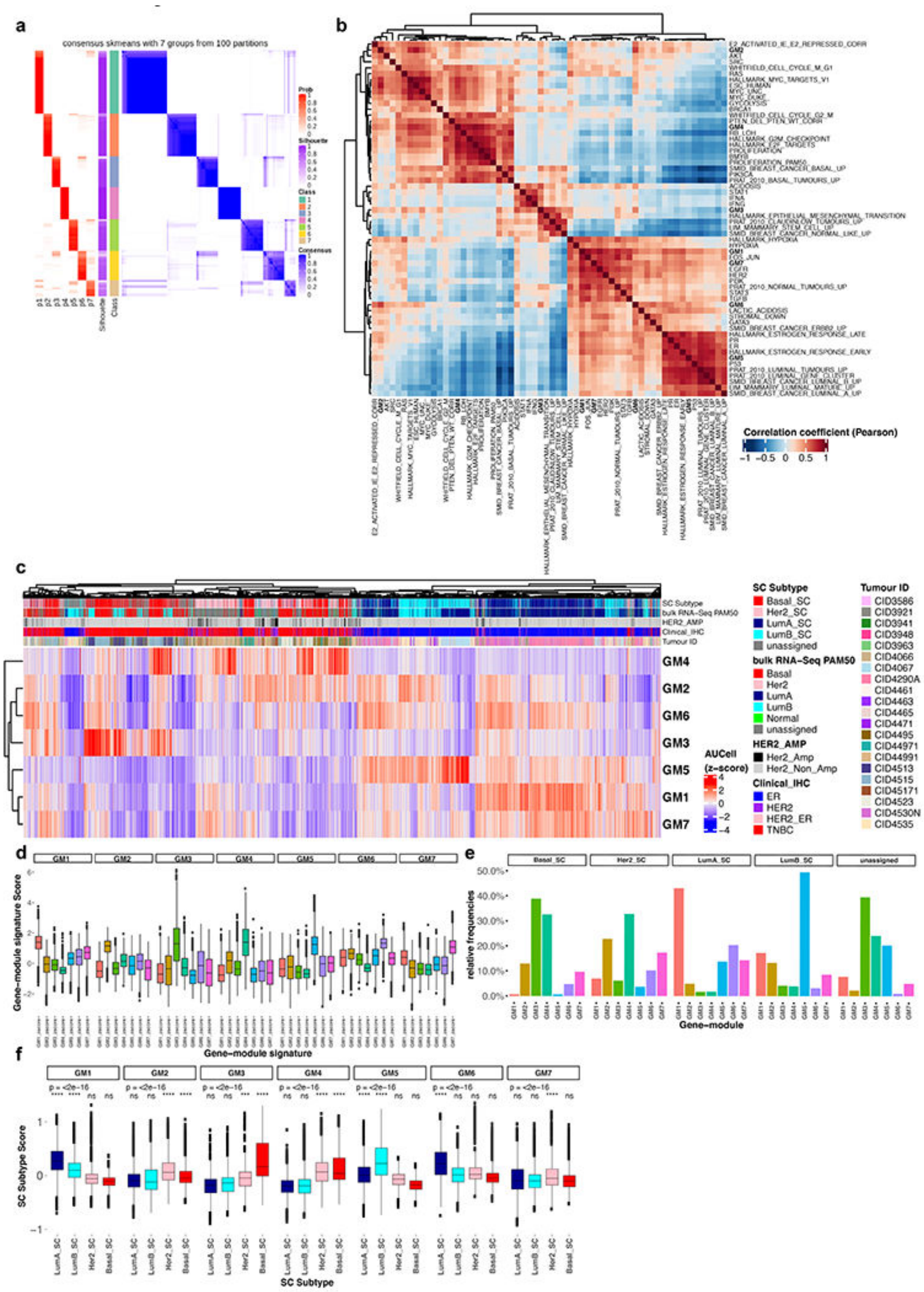
Extended Data Fig. 1. Identification of malignant cells, ***single-cell RNA sequencing metrics and non-integrated data of stromal and immune cells***

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).



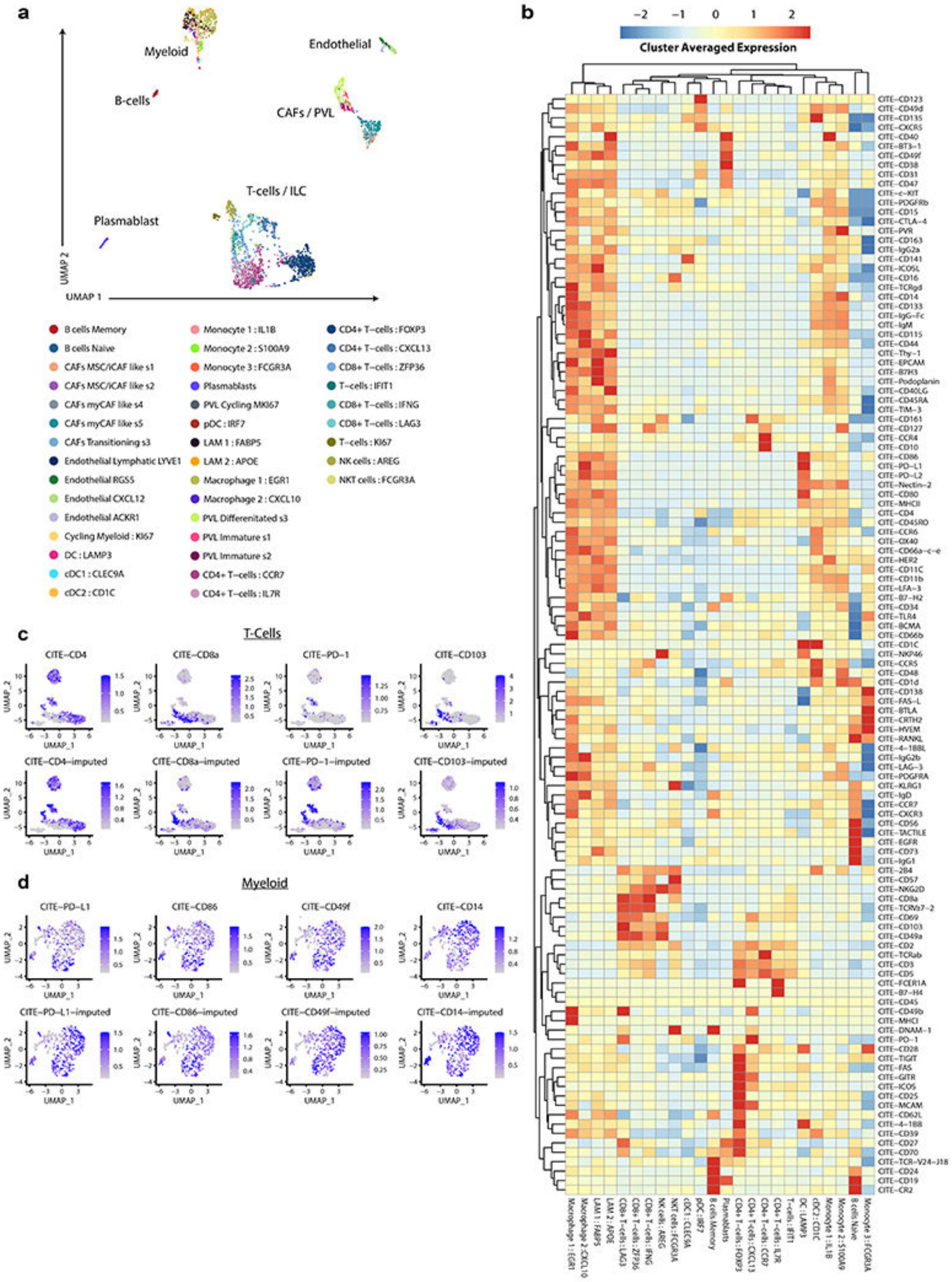
Extended Data Fig. 2. *Supplementary data for* scSubtype classifier

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.



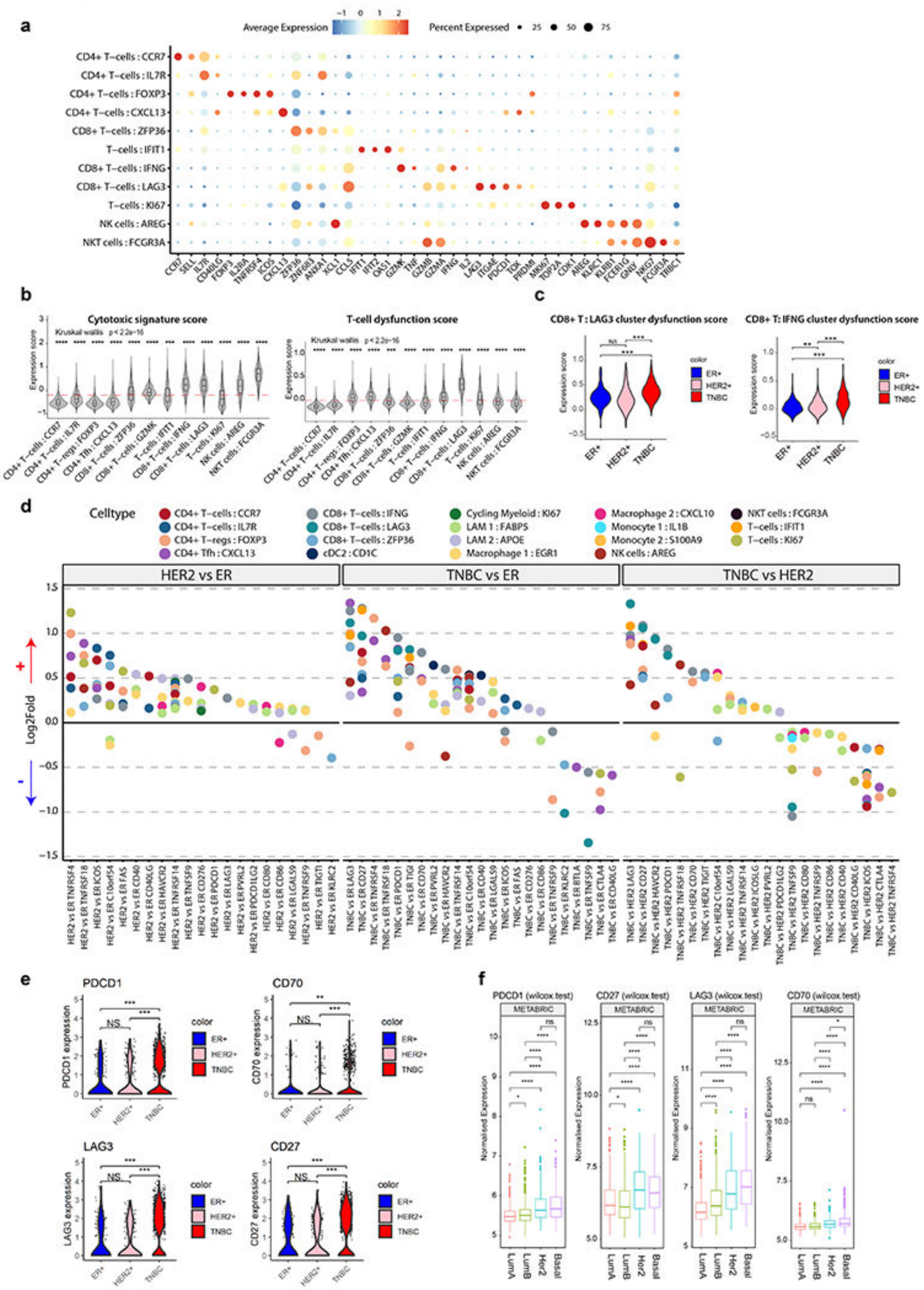
Extended Data Fig. 3. Supplementary data for breast cancer gene modules

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.



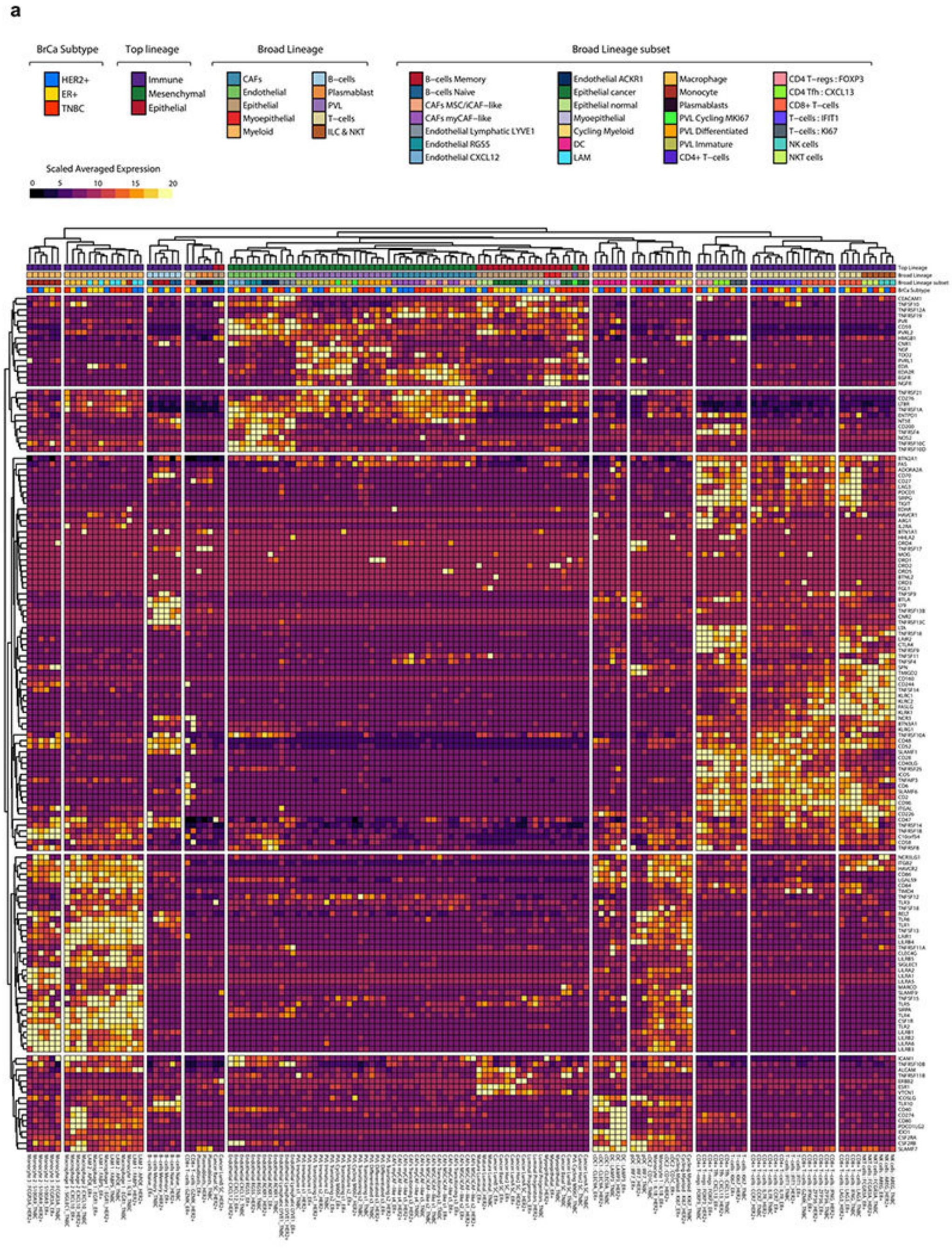
Extended Data Fig. 4. CITE-Seq vignette

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.



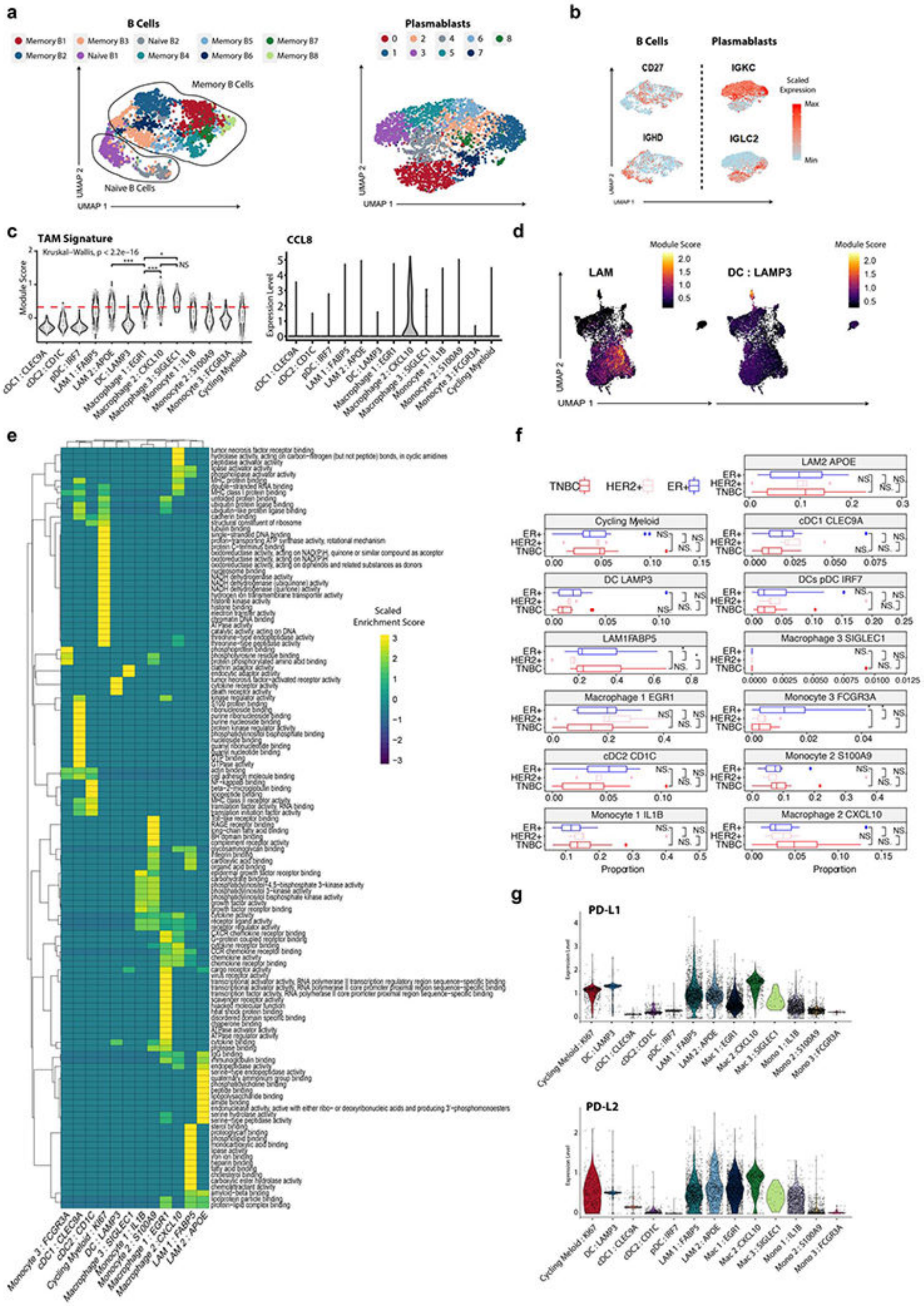
Extended Data Fig. 5. Supplementary data for T-cells and innate lymphoid cells.

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), CD70 and CD27 Log-normalised expression found in LAG3/c8 T-cells across breast cancer subtypes (n = 26; 11 TNBC, 10 ER+ and 5 HER2+). **f,** Enrichment of PDCD1, CD27, LAG3, CD70 expression in 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. Gene expression of immune cell surface receptors across malignant, immune and mesenchymal clusters and breast cancer clinical subtypes

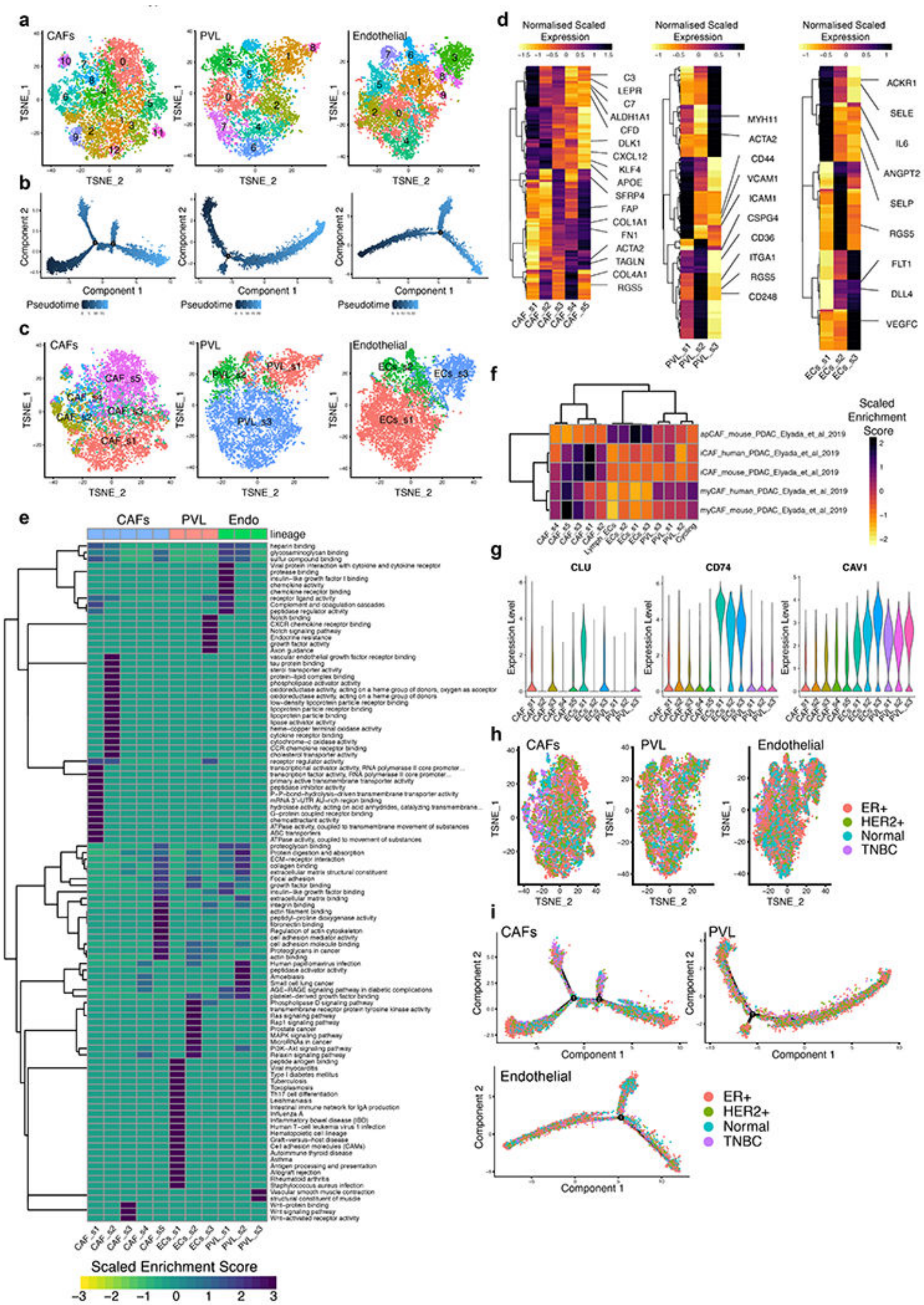
a, Averaged expression and clustering of 133 clinically targetable receptor or ligand immune modulator markers across all cell types grouped by clinical breast cancer subtypes (TNBC, HER2+ and ER+). Gene list was manually curated through systematic literature search of known immune modulating proteins 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. 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,

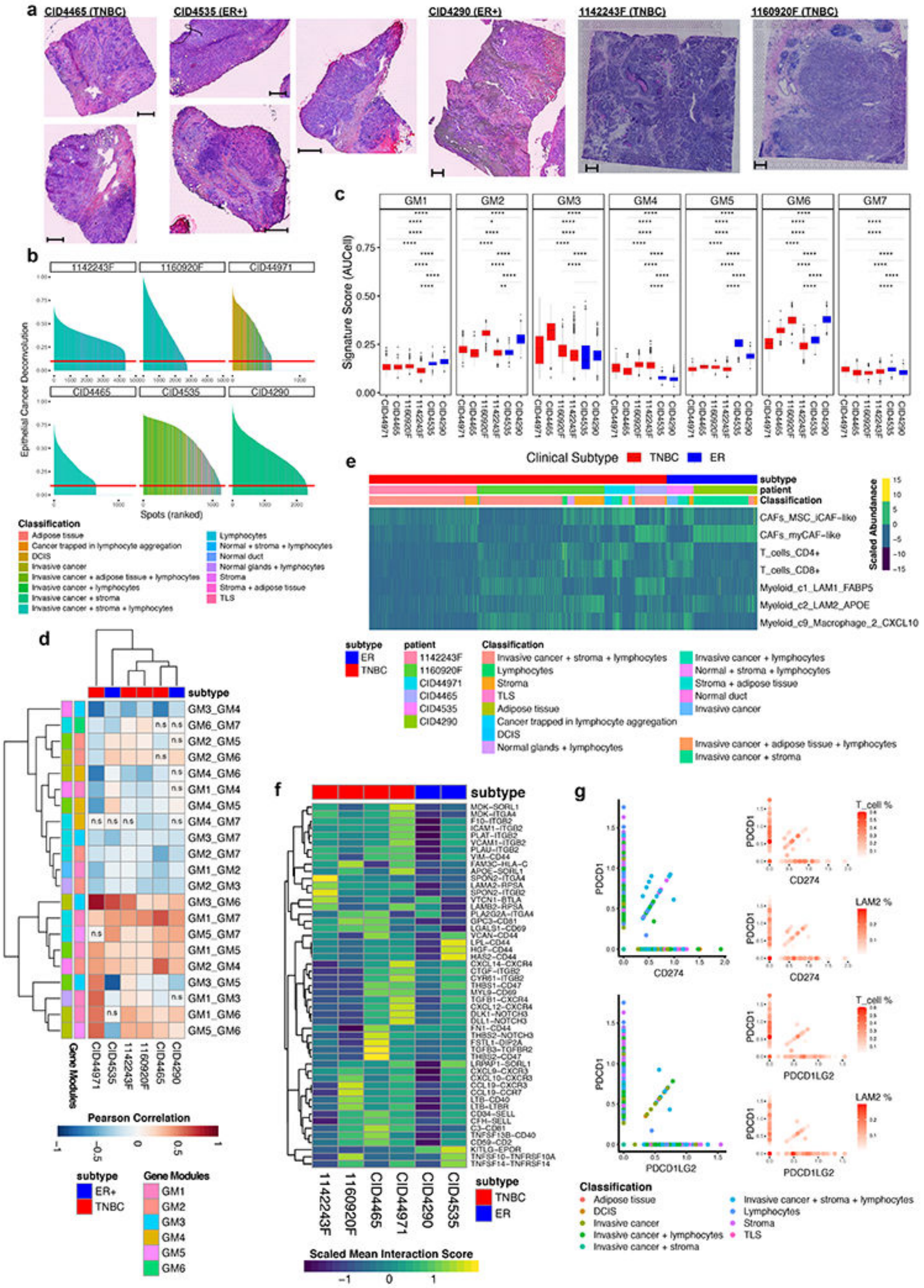
***p < 0.001 and ****p < 0.0001. Dashed red line marks median TAM module score or gene expression. A Kruskal-Wallis test was performed to compare significance between groups'. **d,** LAM and DC : LAMP3 gene expression signatures acquired from Jaitin et al. 2019 and Zhang et al. 2019 respectively, visualized on UMAP myeloid clusters. **e,** Heatmap visualizing GO enrichment pathways across Myeloid clusters. **f,** Proportional 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 plot 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. Supplementary data for mesenchymal cell states and subclusters

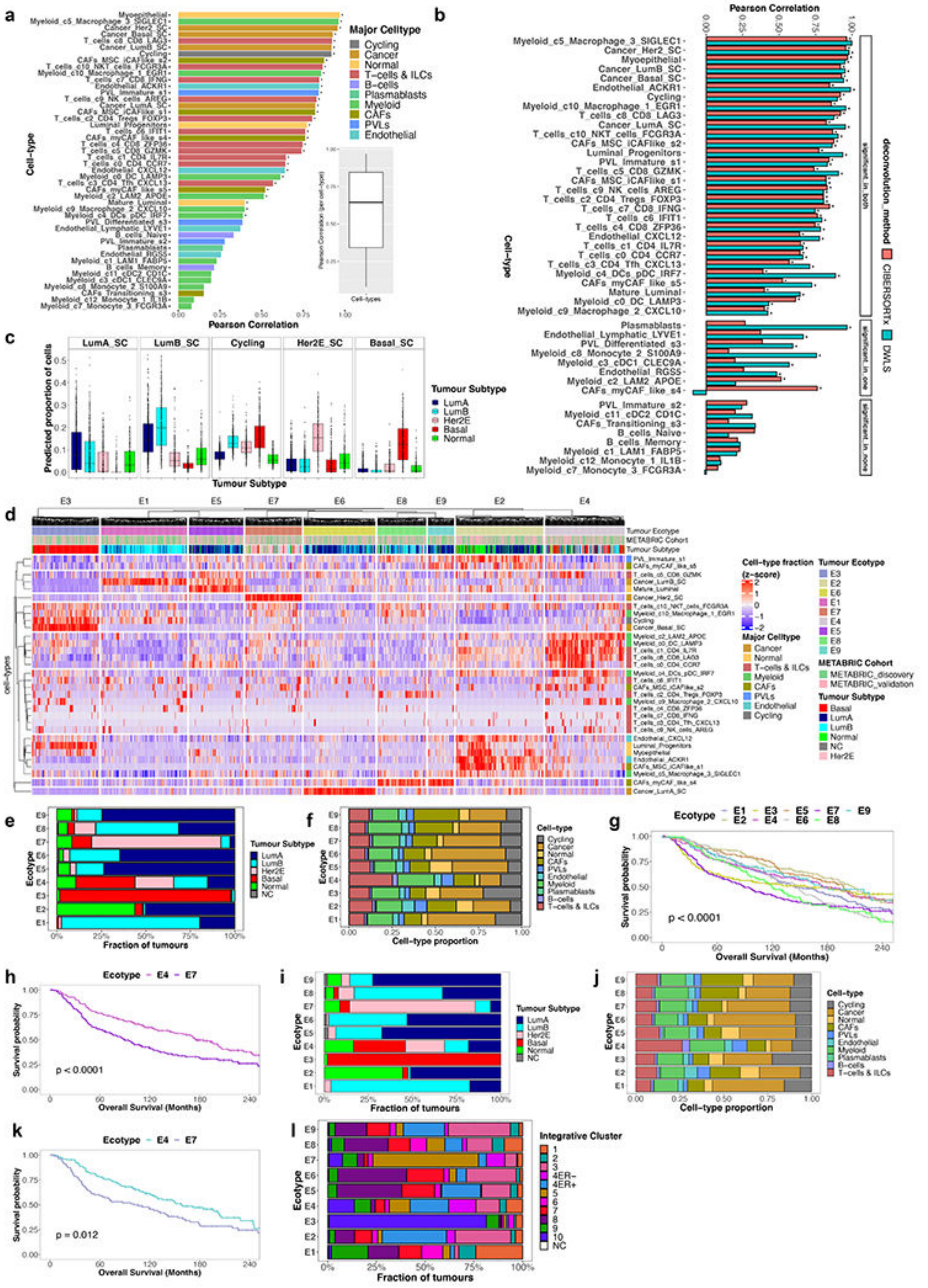
a, UMAP visualization CAFs, PVL cells and endothelial cells using Seurat reclustered with default resolution parameters (0.8). **b,** Pseudotime plot for CAFs, PVL cells and endothelial cells, as determined using monocle. Coordinates are as in main Figure 5c, 5e and 5g. **c,** UMAP visualizations for 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. 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 plot 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. 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work is supported by a research grant from The National Breast Cancer Foundation (NBCF) of Australia (IIRS-19-106) and supported by the generosity of John McMurtrie (AM) and Deborah McMurtrie, The Petre Foundation, The White Butterfly Foundation, The Sydney Breast Cancer Foundation, Skipper Jacobs Charitable Trust, GP Harris Foundation and the NHMRC-funded Centre for Translational Breast Cancer Research. AS is the recipient of a Senior Research Fellowship from the National Health and Medical Research Council of Australia. SZW, GA-E and JT are supported by the Australian Government Research Training Program Scholarship. SOT is supported by the NBCF (PRAC 16-006; IIRS-19-084), The Sydney Breast Cancer Foundation and the Family and Friends of Michael O'Sullivan. SJ is supported by a research fellowship from the NBCF. XSL is supported by the Breast Cancer Research Foundation (BCRF-19-100) and the National Institute of Health of the United States (R01CA234018). CMP and AT were supported by NCI Breast SPORE program (P50-CA58223), RO1-CA148761, and Breast Cancer Research Foundation. This work was supported by the Australian Centre for Translational Breast Cancer Research (TransBCR), Walter and Eliza Hall Institute, with funding support from the NHMRC Centre for Research Excellence grant APP1153049. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

We would like to thanks the following people for their assistance in the experimental part of this manuscript; Ms. Gillian Lehrbach from the Garvan Tissue Culture Facility; Ms. Anaiis Zaratzian from the Garvan Histopathology Facility for tissue processing and IHC staining and guidance on Visium experiments; The Garvan-Weizmann Centre for Cellular Genomics, including Mr. Eric Lam, Ms. Hira Saeed and Ms. Melissa Armstrong for the expertise in flow sorting. We would like to thank Holly Holliday for the incredible illustration in Figure 8g. We thank Heloisa Helena Milioli for providing guidance for analysing the METABRIC cohort dataset. We acknowledge the contribution of Isabelle Shapiro and Cheryl Grant as consumer advocates. This manuscript was edited at Life Science Editors.

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. Processed scRNA-Seq data from this study is also available through the GEO Series accession number GSE176078. 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). ST data from the Andersson *et al.* study[68] can be downloaded from the Zenodo data repository (DOI: 10.5281/zenodo.3957257).

References

1. Kim HK 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). [PubMed: 30189722]
2. Picornell AC 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). [PubMed: 31159741]
3. Parker JS et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 27, 1160–7 (2009). [PubMed: 19204204]
4. Perou CM et al. Molecular portraits of human breast tumours. Nature 406, 747–52 (2000). [PubMed: 10963602]
5. Sorlie T et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 98, 10869–74 (2001). [PubMed: 11553815]
6. Sorlie T et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A 100, 8418–23 (2003). [PubMed: 12829800]
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). [PubMed: 29395328]
8. Cazet AS et al. Targeting stromal remodeling and cancer stem cell plasticity overcomes chemoresistance in triple negative breast cancer. Nature Communications 9, 2897 (2018).
9. Dushyanthen S et al. Relevance of tumor-infiltrating lymphocytes in breast cancer. BMC Med 13, 202 (2015). [PubMed: 26300242]
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). [PubMed: 30930117]
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). [PubMed: 32783915]
12. Medler TR et al. Complement C5a Fosters Squamous Carcinogenesis and Limits T Cell Response to Chemotherapy. Cancer Cell 34, 561–578 e6 (2018). [PubMed: 30300579]
13. Nakamura K & Smyth MJ TREM2 marks tumor-associated macrophages. Signal Transduct Target Ther 5, 233 (2020). [PubMed: 33037186]
14. Costa A et al. Fibroblast Heterogeneity and Immunosuppressive Environment in Human Breast Cancer. Cancer Cell 33, 463–479 e10 (2018). [PubMed: 29455927]
15. Wu SZ et al. Stromal cell diversity associated with immune evasion in human triple-negative breast cancer. EMBO J, e104063 (2020). [PubMed: 32790115]
16. Sahai E et al. A framework for advancing our understanding of cancer-associated fibroblasts. Nat Rev Cancer 20, 174–186 (2020). [PubMed: 31980749]
17. Ohlund D et al. Distinct populations of inflammatory fibroblasts and myofibroblasts in pancreatic cancer. J Exp Med 214, 579–596 (2017). [PubMed: 28232471]
18. Biffi G et al. IL1-Induced JAK/STAT Signaling Is Antagonized by TGFbeta to Shape CAF Heterogeneity in Pancreatic Ductal Adenocarcinoma. Cancer Discov (2018).

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). [PubMed: 31197017]

20. Puram SV et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell 171, 1611–1624 e24 (2017). [PubMed: 29198524]

21. Lambrechts D et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med 24, 1277–1289 (2018). [PubMed: 29988129]

22. Azizi E et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell 174, 1293–1308 e36 (2018). [PubMed: 29961579]

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). [PubMed: 29942092]

24. Kim C et al. Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. Cell 173, 879–893 e13 (2018). [PubMed: 29681456]

25. Ali HR et al. Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. Nature Cancer 1, 163–175 (2020). [PubMed: 35122013]

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). [PubMed: 30982598]

27. Aran D, Hu Z & Butte AJ xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol 18, 220 (2017). [PubMed: 29141660]

28. Lim E et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nat Med 15, 907–13 (2009). [PubMed: 19648928]

29. Pliner HA, Shendure J & Trapnell C Supervised classification enables rapid annotation of cell atlases. Nature Methods 16, 983–986 (2019). [PubMed: 31501545]

30. Neftel C et al. An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell 178, 835–849 e21 (2019). [PubMed: 31327527]

31. Tirosh I et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352, 189–96 (2016). [PubMed: 27124452]

32. Cancer Genome Atlas N Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70 (2012). [PubMed: 23000897]

33. Prat A et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res 12, R68 (2010). [PubMed: 20813035]

34. Nielsen TO 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, 5222–32 (2010). [PubMed: 20837693]

35. Stoeckius M et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods 14, 865–868 (2017). [PubMed: 28759029]

36. Stuart T et al. Comprehensive Integration of Single-Cell Data. Cell 177, 1888–1902 e21 (2019). [PubMed: 31178118]

37. Glajcar A, Szpor J, Hodorowicz-Zaniewska D, Tyrak KE & Okon 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). [PubMed: 31016433]

38. Li H et al. Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. Cell 176, 775–789 e18 (2019). [PubMed: 30595452]

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, 424–30 (2002). [PubMed: 12452832]

40. Curtis C et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486, 346–52 (2012). [PubMed: 22522925]

41. Slyper M et al. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. Nature Medicine 26, 792–802 (2020).

42. Ruffell B et al. Leukocyte composition of human breast cancer. Proc Natl Acad Sci U S A 109, 2796–801 (2012). [PubMed: 21825174]

43. Zhang Q et al. Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma. Cell 179, 829–845 e20 (2019). [PubMed: 31675496]

44. Jaitin DA et al. Lipid-Associated Macrophages Control Metabolic Homeostasis in a Trem2-Dependent Manner. Cell 178, 686–698 e14 (2019). [PubMed: 31257031]

45. Chen J et al. CCL18 from tumor-associated macrophages promotes breast cancer metastasis via PITPNM3. Cancer Cell 19, 541–55 (2011). [PubMed: 21481794]

46. Qiu X et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods 14, 979–982 (2017). [PubMed: 28825705]

47. Kumar A et al. Specification and Diversification of Pericytes and Smooth Muscle Cells from Mesenchymoangioblasts. Cell Rep 19, 1902–1916 (2017). [PubMed: 28564607]

48. Thiriot A et al. Differential DARC/ACKR1 expression distinguishes venular from non-venular endothelial cells in murine tissues. BMC Biology 15, 45 (2017). [PubMed: 28526034]

49. Mailhos C et al. Delta4, an endothelial specific notch ligand expressed at sites of physiological and tumor angiogenesis. Differentiation 69, 135–44 (2001). [PubMed: 11798067]

50. Ubezio B et al. Synchronization of endothelial Dll4-Notch dynamics switch blood vessels from branching to expansion. Elife 5(2016).

51. Kryczek I et al. CXCL12 and vascular endothelial growth factor synergistically induce neoangiogenesis in human ovarian cancers. Cancer Res 65, 465–72 (2005). [PubMed: 15695388]

52. Blanco R & Gerhardt H VEGF and Notch in tip and stalk cell selection. Cold Spring Harb Perspect Med 3, a006569 (2013). [PubMed: 23085847]

53. Jakobsson L et al. Endothelial cells dynamically compete for the tip cell position during angiogenic sprouting. Nat Cell Biol 12, 943–53 (2010). [PubMed: 20871601]

54. Andersson A et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. Communications Biology 3, 565 (2020). [PubMed: 33037292]

55. Lakins MA, Ghorani E, Munir H, Martins CP & Shields JD Cancer-associated fibroblasts induce antigen-specific deletion of CD8 (+) T Cells to protect tumour cells. Nat Commun 9, 948 (2018). [PubMed: 29507342]

56. Newman AM et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnology 37, 773–782 (2019).

57. Tsoucas D et al. Accurate estimation of cell-type composition from gene expression data. Nat Commun 10, 2975 (2019). [PubMed: 31278265]

58. Tsai JH, Donaher JL, Murphy DA, Chau S & Yang J Spatiotemporal regulation of epithelial-mesenchymal transition is essential for squamous cell carcinoma metastasis. Cancer Cell 22, 725–36 (2012). [PubMed: 23201165]

59. Katzenelenbogen Y et al. Coupled scRNA-Seq and Intracellular Protein Activity Reveal an Immunosuppressive Role of TREM2 in Cancer. Cell (2020).

60. Molgora M et al. TREM2 Modulation Remodels the Tumor Myeloid Landscape Enhancing Anti-PD-1 Immunotherapy. Cell (2020).

61. Wu SZ et al. Cryopreservation of human cancers conserves tumour heterogeneity for single-cell multi-omics analysis. Genome Medicine 13, 81 (2021). [PubMed: 33971952]

62. Lun ATL et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol 20, 63 (2019). [PubMed: 30902100]

63. Zhao X, Rodland EA, Tibshirani R & Plevritis S Molecular subtyping for clinically defined breast cancer subgroups. Breast Cancer Res 17, 29 (2015). [PubMed: 25849221]

64. Patro R, Duggal G, Love MI, Irizarry RA & Kingsford C Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods 14, 417–419 (2017). [PubMed: 28263959]

65. 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). [PubMed: 26653891]

66. Aibar S et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods 14, 1083–1086 (2017). [PubMed: 28991892]

67. Yu G, Wang LG, Han Y & He QY clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16, 284–7 (2012). [PubMed: 22455463]

68. Andersson A et al. Spatial Deconvolution of HER2-positive Breast Tumors Reveals Novel Intercellular Relationships. bioRxiv, 2020.07.14.200600 (2020).

69. Efremova M, Vento-Tormo M, Teichmann SA & Vento-Tormo R CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nature Protocols 15, 1484–1506 (2020). [PubMed: 32103204]

70. Ramilowski JA et al. A draft network of ligand-receptor-mediated multicellular signalling in human. Nat Commun 6, 7866 (2015). [PubMed: 26198319]

71. Wu SZ, Al-Eryani G, Roden D, Bartonicek N & Swarbrick A (2021). BrCa_cell_atlas: Release (Version 1.0.0) [Analysis Code]. Zenodo. DOI: 10.5281/zenodo.5031502

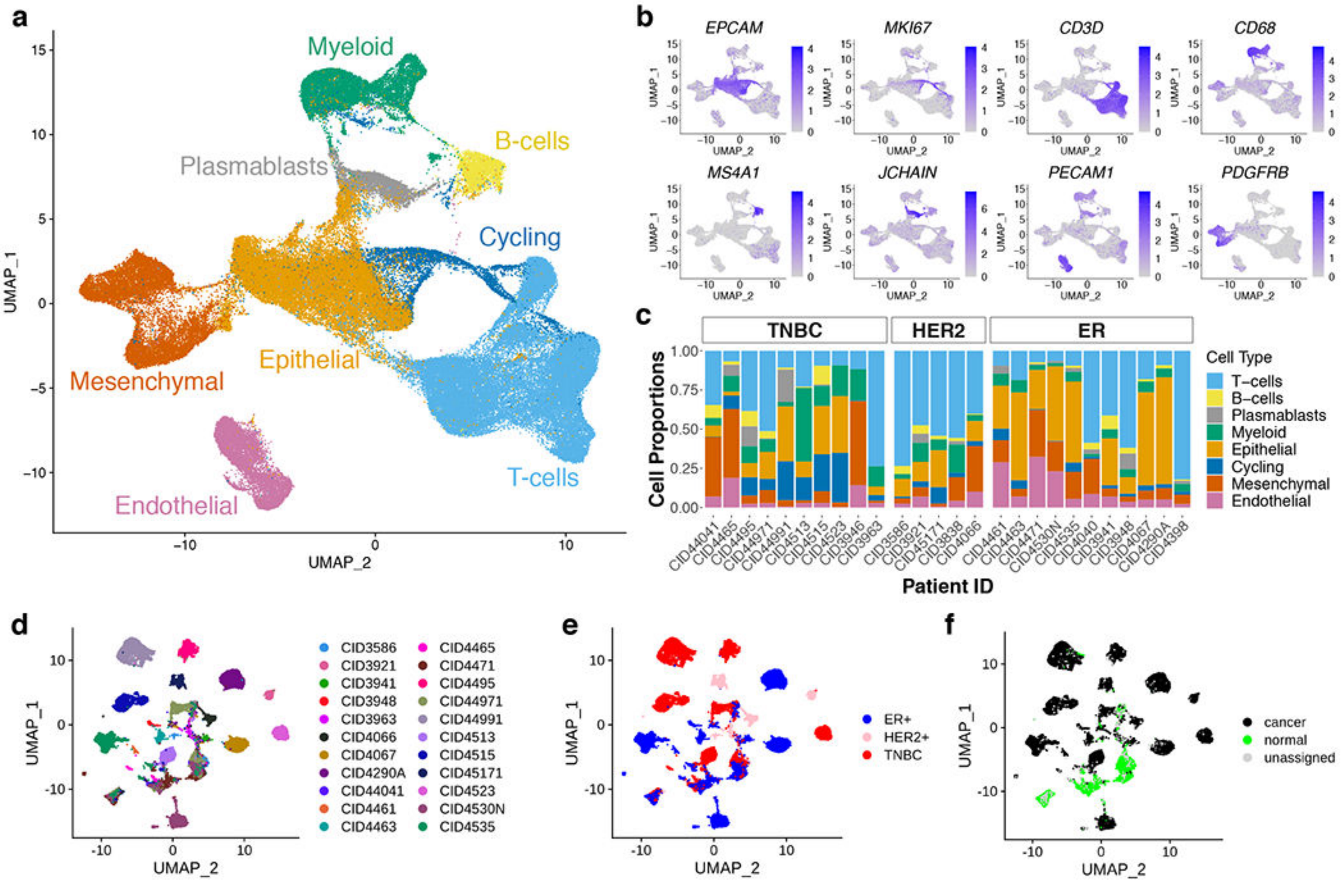


Figure 1.

Cellular composition of primary breast cancers and the identification of malignant epithelial cells. **a,** Integrated dataset overview of 130,246 cells analyzed by scRNA-Seq. Clusters are 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; *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 tumours with at least 200 epithelial cells, colored by tumor (d), clinical subtype (e) and inferCNV classification (f).

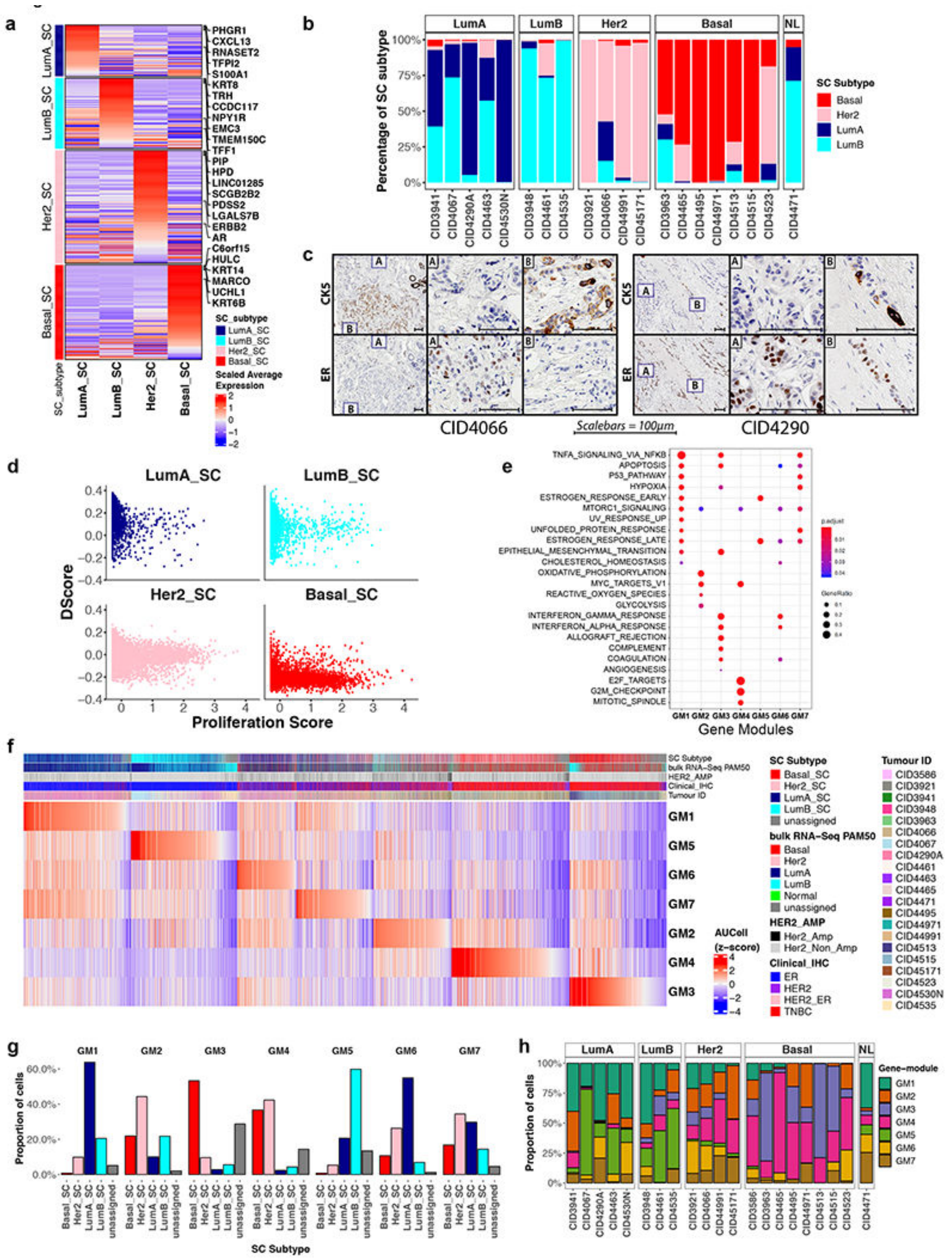


Figure 2.

Identifying drivers of neoplastic breast cancer cell heterogeneity. **a,** Heatmap showing the average expression (scaled) of all cells assigned to each of the four scSubtypes. The top-5 most highly expressed genes in each subtype are shown, and 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) immunohistochemistry (IHC) from two tumors (CID4066, left; CID4290, right) with intrinsic subtype heterogeneity from b (n = 24 breast tumors analysed). The left panel represents whole tissue sections, with two regions of interest labelled (A and B). The middle panel represents CK5-/ER+ areas (insert A), whilst the right panel shows CK5+/ER− areas (insert B). Scale bar represent 100 μm. **d,** Scatter plot of the proliferation scores 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 (Wilcox test p<0.001) for both proliferation and DScores. **e,** Gene-set enrichment, using ClusterProfiler, of the 200 genes in each of the gene-modules (GM1-7). Significantly enriched (bonferroni adjusted p-value < 0.05) gene-sets from the MSigDB HALLMARK collection are shown. **f,** Proportion of cells assigned to each of the scSubtype subtypes grouped according to gene-module. **g,** Scaled signature scores of each of the seven intra-tumor transcriptional heterogeneity gene-modules (rows) across all individual neoplastic cells (columns). Cells are ordered based on the strength of the gene-module signature score. **h,** Percentage of neoplastic cells assigned to each of the seven gene-modules.

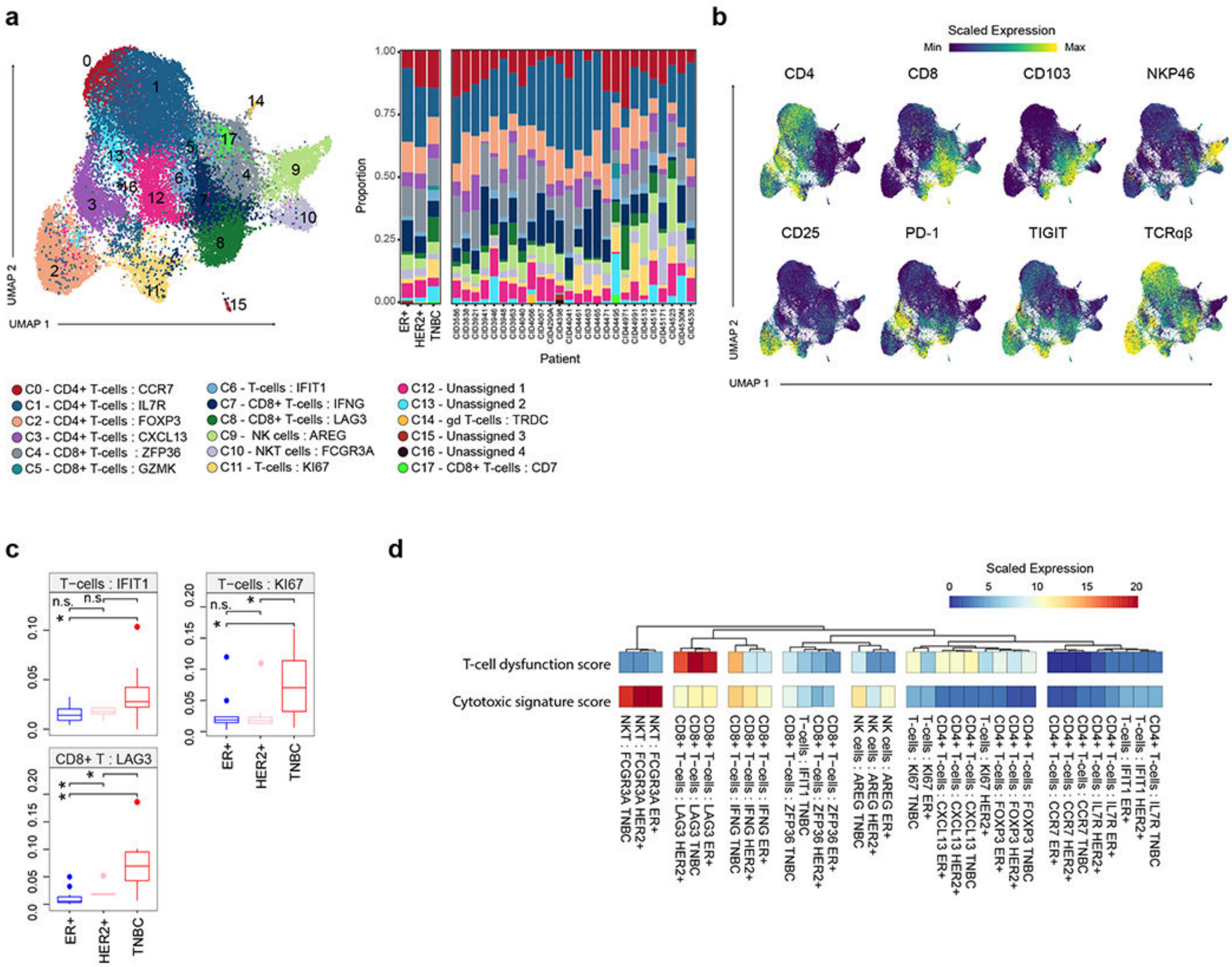


Figure 3.

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, CD8+ T-cells : LAG3 in TNBC tumors, (n = 26; 11 TNBC, 10 ER+ and 5 HER2+). 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. 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-values denoted by asterisks: *p < 0.05, p < 0.01, *p < 0.001 and ****p < 0.0001. **d**, Cluster averaged dysfunctional and cytotoxic effector gene signature scores in T-cells and innate lymphoid cells stratified by clinical subtypes.

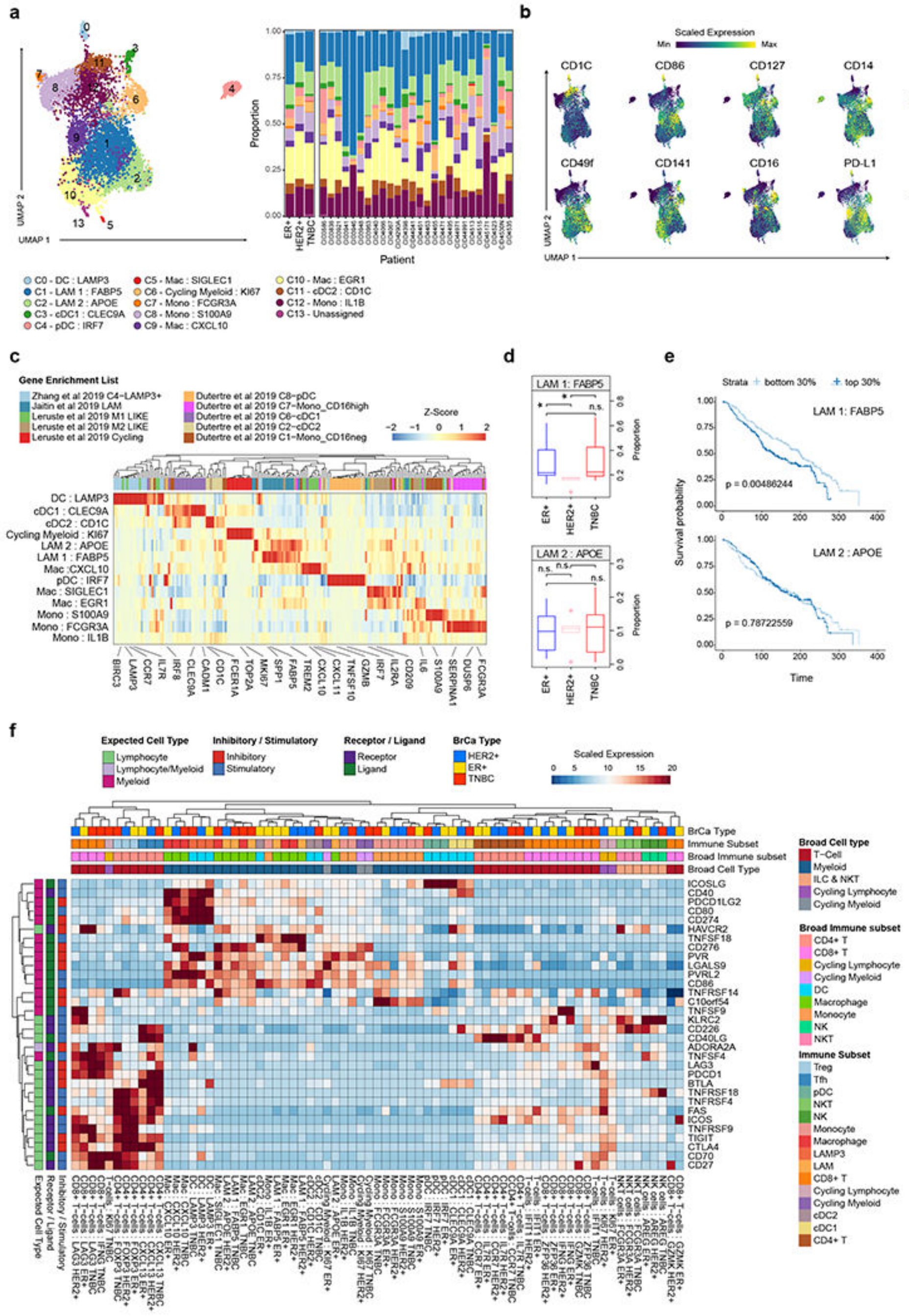


Figure 4.

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. **d,** Proportions of LAM 1 : *FABP5* and LAM 2: *APOE* (n = 26; 11 TNBC, 10 ER+ and 5 HER2+) across clinical subtypes. Box plot 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. 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-values denoted by asterisks: *$p < 0.05$, $p < 0.01$, *$p < 0.001$ and ****$p < 0.0001$. **e,** Kaplan Meier plots showing associations between LAM 1 : FABP5 or LAM 2 : APOE with overall survival in METABRIC cohort (top 30% and bottom 30%, $n = 180$ per group). P-values were calculated using log-rank test. Time (x-axis) is represented in months. **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), the inhibitory (red) or stimulatory status (blue) and the expected major lineage cell types known to express the gene (lymphocyte, green; myeloid, pink; both, light purple).

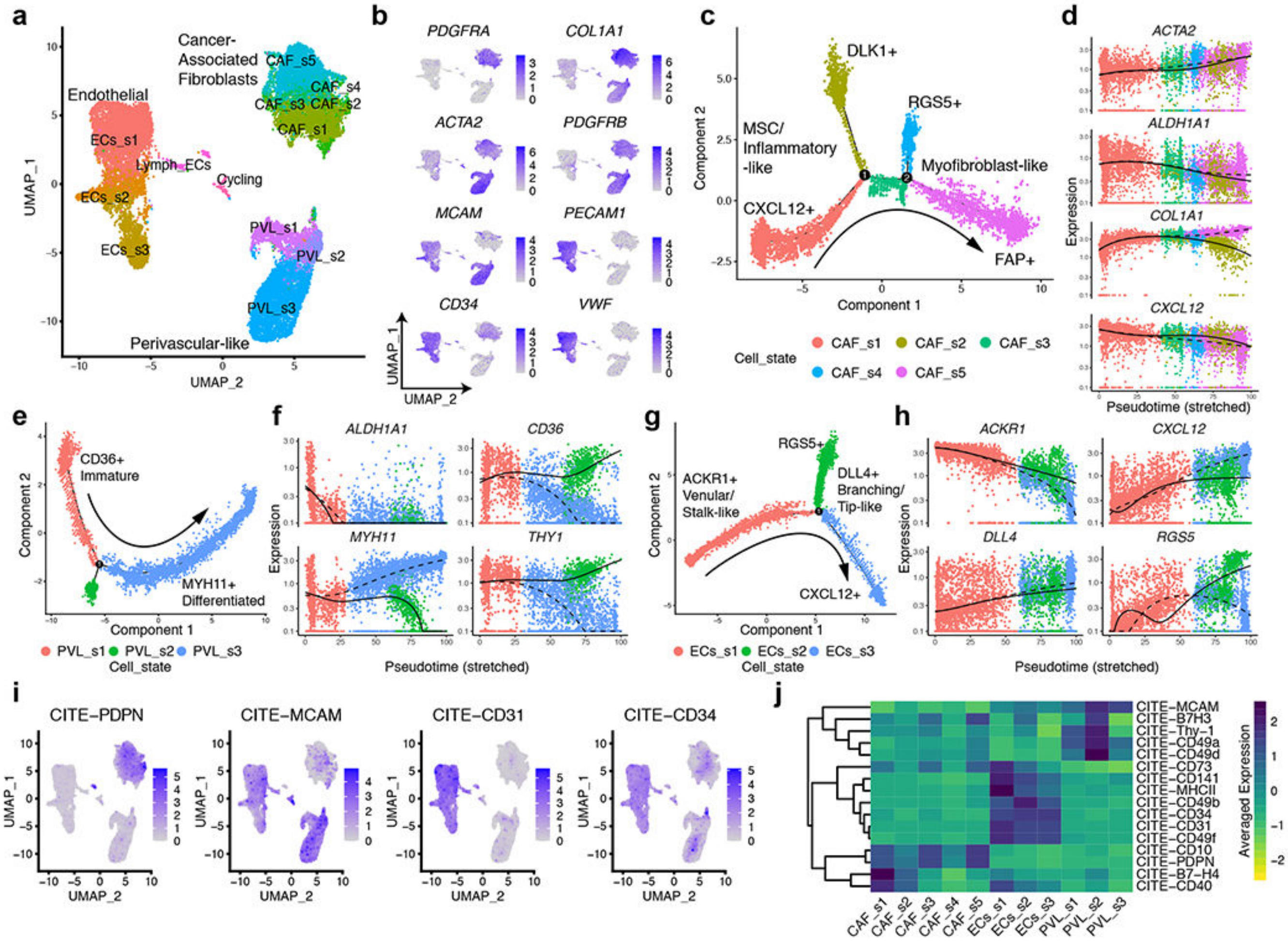


Figure 5.

Transcriptional profiling and phenotyping of diverse mesenchymal differentiation states across breast cancers. **a,** UMAP visualization of reclustered mesenchymal cells, including CAFs (6,573 cells), perivascular-like (PVL) cells (5,423 cells), endothelial cells (7,899 cells; ECs), lymphatic ECs (203 cells) and cycling PVL (50 cells). Cell sub-states are defined using pseudotemporal ordering using Monocle (as in c-h). **b,** Featureplots of canonical markers for CAFs (*PDGFRA, COL1A1, ACTA2, PDGFRB*), PVL (*ACTA2, PDGFRB* and *MCAM*) and ECs (*PECAM1, CD34* and *VWF*). UMAP axes correspond to Figure 5a. 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 ECs (g-h). **c-d,** Five states of CAFs: CAF s1 and s2 both resemble mesenchymal stem cells (MSC; *ALDH1A1*) and inflammatory CAF-like states (iCAF; *CXCL12*); CAF s2 was distinct from s1 by DLK1; CAF s4 and s5 resemble myofibroblast-like states (myCAF; *ACTA2*) which were enriched for ECM genes (*COL1A1*); transitioning CAF s3 shared features of both MSC/iCAFs and myCAFs. **e-f,** Three PVL states: s1 and s2 resemble progenitor and immature states (imPVL; *ALDH1A1*); PVL s3 resembles a contractile and differentiated state (dPVL; *MYH11*). **g-h,** Three EC states: s1 resemble a venular stalk-like state (*ACKR1*) and two tip-like states (*DLL4*), s2 and s3, that are distinguished by *RGS5* and *CXCL12*, respectively. **i,** Featureplots of imputed CITE-Seq antibody-derived tag (ADT) protein levels for canonical markers of CAFs (PDPN), PVL cells (CD146/*MCAM*) and ECs (CD31 and CD34). UMAP coordinates correspond to those in a. **j,** Heatmap of cluster averaged imputed CITE-Seq values for additional cell surface markers and functional molecules.

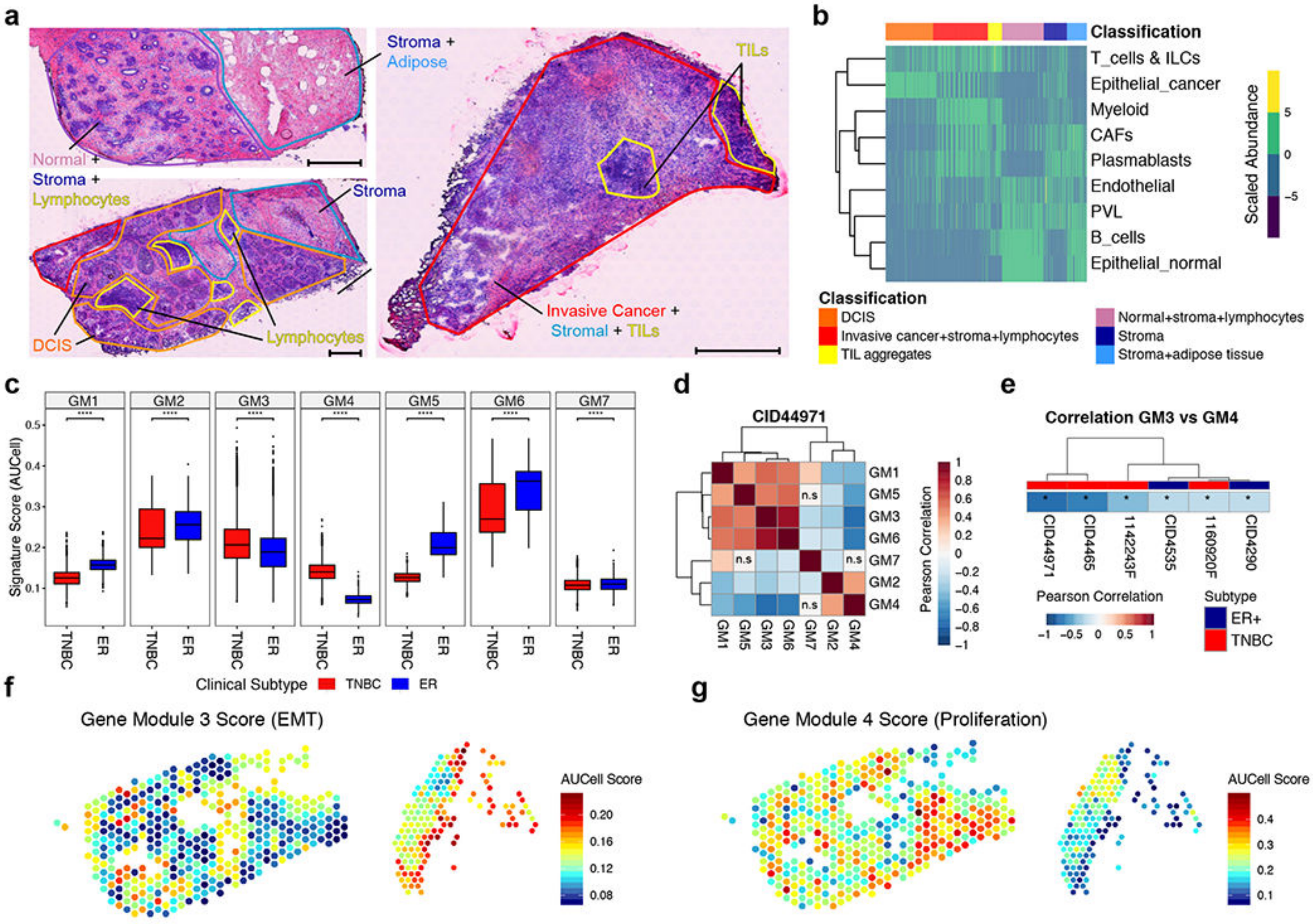


Figure 6.

Mapping breast cancer heterogeneity using spatial transcriptomics. **a,** Complete H&E images of all three tissue regions analysed using Visium for the sample TNBC CID44971. Pathological annotation of morphological regions into distinct categories including normal ductal (green), stroma and adipose (blue), lymphocyte aggregates (yellow), ductal carcinoma in-situ (DCIS; orange) and invasive cancer (red). Black scale bars represent 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 gene module scores grouped by clinical subtype across the six cases (n = 11,535 spots from 4 x TNBC tumors and 2 x 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. 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,** Pearson correlation heatmap of breast cancer gene modules in TNBC CID44971 (“n.s” represent non-significant correlations; two-sided correlation coefficient, Benjamini–Hochberg adjusted p-values < 0.05). Spots with high cancer epithelial abundances (>10%), were scored with gene module

(GM) signatures using AUCell. **e,** Negative correlation between GM3 (EMT) and GM4 (Proliferation/Cell Cycle) across all cancer epithelial spots from six breast cancers analysed by ST (two-sided correlation coefficient, *denotes p-value < 0.05). **f-g,** Scaled 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 Figure 6a.

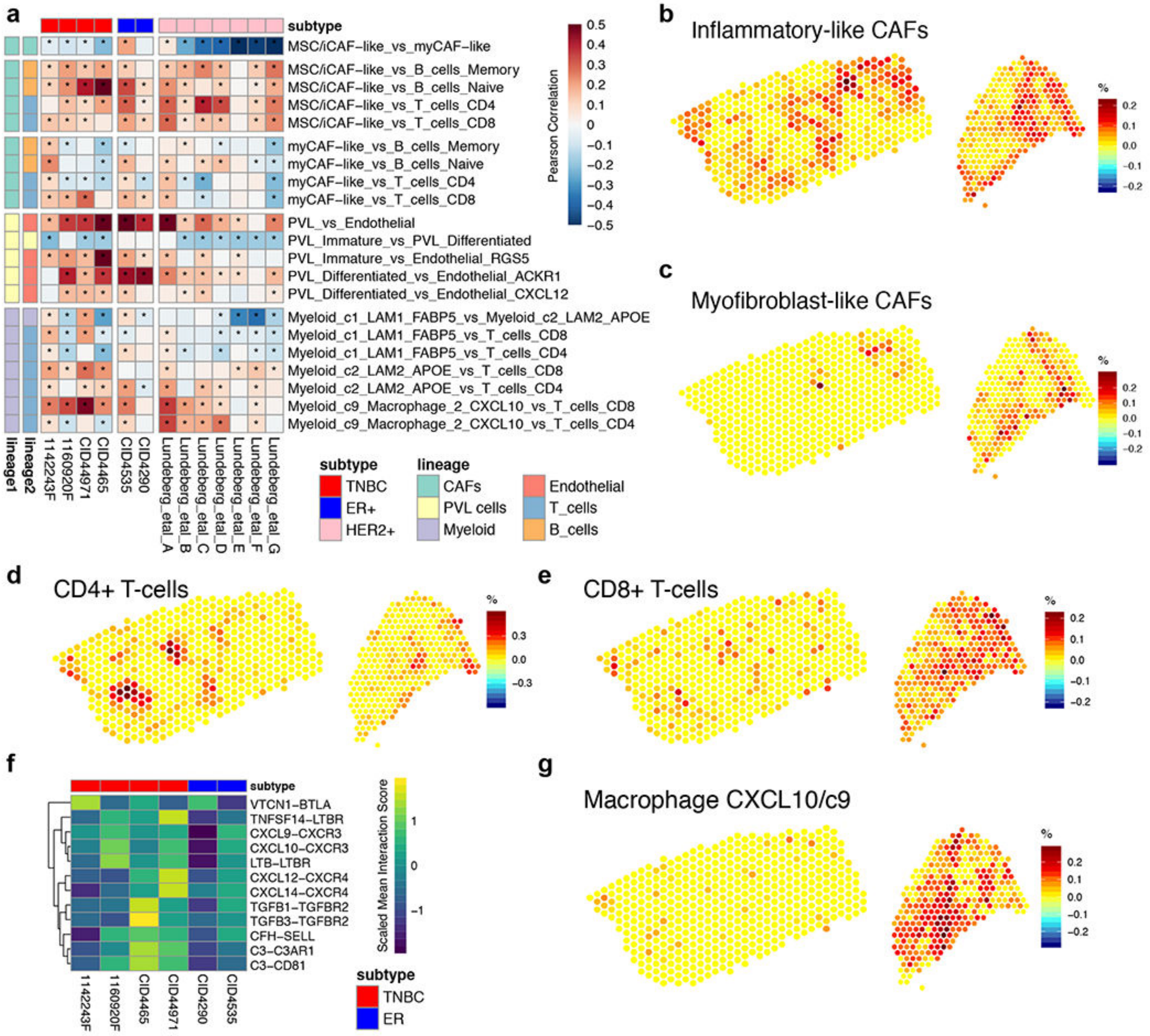


Figure 7.

Spatially mapping novel 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, *denotes Benjamini–Hochberg adjusted p-value < 0.05). Each tumor is stratified by the clinical subtype, including four TNBC (blue) and two ER+ (red) analysed in this study and seven HER2+ (pink) cases from the Lundeberg et al. study. **b-e,** Scaled deconvolution values for iCAFs (b), myCAFs (c), CD4+ (d) and CD8+ T-cells (e) overlaid onto tissue spots, as defined in the bottom left and right tissue sections in Figure 6a. Representative TNBC case CID44971 is shown. **f,** Spatial proximity of selected CAF T-cell signalling 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 Figure 6a. Representative TNBC case CID44971 is shown.

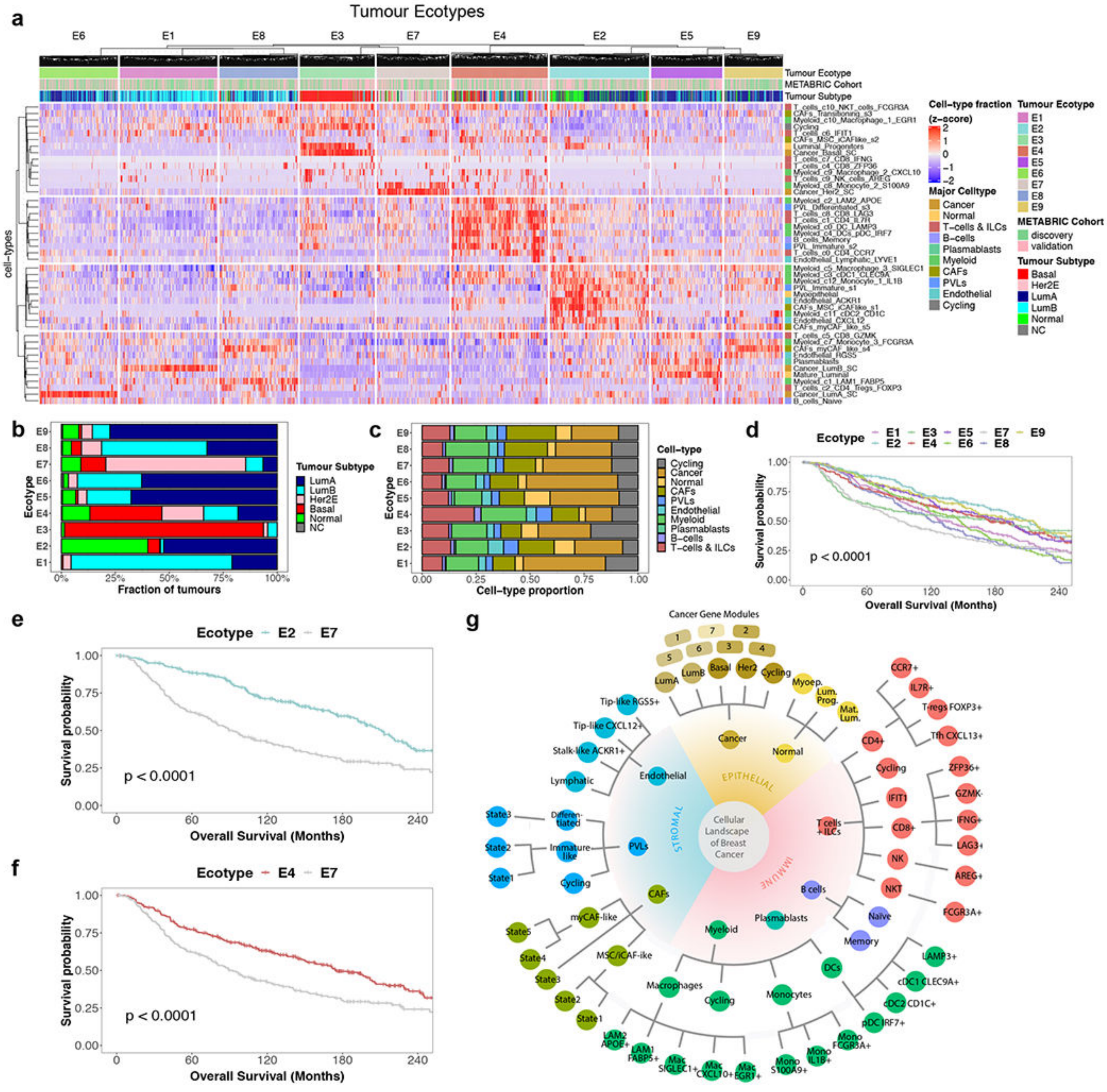


Figure 8.

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 METABRIC showing nine 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, 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 (KM) plot of the patients with tumors in each of the nine ecotypes (d), patients with tumors in ecotypes E2 and E7 (e), patients with tumors in ecotypes E4 and E7 (f). p-values calculated using the log-rank test. **g,** Summary of the major epithelial, immune and stromal cell types identified in this study grouped by their major (inner), minor and subset (outer) level classification tiers.