nature cancer

ARTICLES
https://doi.org/10.1038/s43018-020-0026-6

Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer

H. Raza Ali[1,2,6], Hartland W. Jackson[1,6], Vito R. T. Zanotelli[1], Esther Danenberg[1], Jana R. Fischer[1], Helen Bardwell[2], Elena Provenzano[3], CRUK IMAXT Grand Challenge Team[4], Oscar M. Rueda[2], Suet-Feung Chin[2], Samuel Aparicio[5], Carlos Caldas[2,3]✉ and Bernd Bodenmiller[1]✉

Genomic alterations shape cell phenotypes and the structure of tumor ecosystems in poorly defined ways. To investigate these relationships, we used imaging mass cytometry to quantify the expression of 37 proteins with subcellular spatial resolution in 483 tumors from the METABRIC cohort. Single-cell analysis revealed cell phenotypes spanning epithelial, stromal and immune types. Distinct combinations of cell phenotypes and cell–cell interactions were associated with genomic subtypes of breast cancer. Epithelial luminal cell phenotypes separated into those predominantly impacted by mutations and those affected by copy number aberrations. Several features of tumor ecosystems, including cellular neighborhoods, were linked to prognosis, illustrating their clinical relevance. In summary, systematic analysis of single-cell phenotypic and spatial correlates of genomic alterations in cancer revealed how genomes shape both the composition and architecture of breast tumor ecosystems and will enable greater understanding of the phenotypic impact of genomic alterations.

The heterogeneity of cancer remains an obstacle to effective clinical management. Efforts to understand this inter-tumor heterogeneity in breast cancer have identified tumor subtypes associated with distinct clinical behaviors[1–3] and driver genomic alterations[4–6]. However, these classifications do not account for the cellular complexity of solid tumors, which comprise diverse cancerous and non-cancerous cells in distinct spatial arrangements and in a variety of transitory states[7]. Genomic alterations within cancer cells likely determine the components and structures of these multicellular ecosystems, which ultimately drive disease progression and treatment resistance. Thus, an understanding of how genomic alterations shape tumor ecosystems should enable identification of biomarkers and development of new treatments. Here we studied, in unprecedented detail, how genomic alterations shape breast tumor ecosystems by coupling imaging mass cytometry[8] (IMC) to multiplatform genomics. We quantified the abundances of 37 markers in 483 breast tumor samples from the METABRIC cohort[2,5,9], enabling a systematic 'phenogenomic' analysis of breast cancer.

Results

Spatially resolved phenotyping of breast tumor ecosystems by IMC. To study the cellular composition of breast tumors while preserving spatial context, we used IMC to detect 37 proteins in formalin-fixed, paraffin-embedded samples of 483 tumors from the METABRIC cohort. These tumors have undergone extensive genomic characterization, including copy number, transcriptomic and microRNA (miRNA) profiling and targeted sequencing of 173 breast cancer-associated genes[2,5,9] (Fig. 1a and Supplementary Table 1). Tissues were stained with a panel of isotope-labeled antibodies (Supplementary Table 2). Stained sections were laser ablated at subcellular resolution, and liberated isotopes were detected with a mass cytometer[8] to yield images revealing the abundance and location of the 37 proteins of interest simultaneously (Fig. 1b).

We analyzed the resulting data by using an image processing pipeline adapted for IMC[10–12]. Briefly, we used random forest classification to segment single cells and then quantified the expression of proteins per cell and recorded the identities of adjacent cells[13]. The resulting multiplexed molecular tissue maps, taken together with extensive matched publicly available genomic data[2,5,9,14], characterized these breast tumors with unprecedented depth, linking multidimensional tumor phenotypes with somatic genomic alterations.

Data-driven derivation of cell phenotypes. To investigate cellular diversity and intercellular relationships in breast tumors, we analyzed IMC-derived single-cell expression data by using a combination of clustering approaches (Fig. 1c). The resulting cell phenotypes fell broadly into the categories of tumor, stromal and immune cells (Fig. 2a,b). Most cells were epithelial (Fig. 2c). We determined cell identities by comparison of lineage marker expression and inspection of cell morphology and location (Fig. 2d–f and Extended Data Fig. 1). There was diversity among cells categorized as fibroblasts or myofibroblasts. Myofibroblasts were distinguished from fibroblasts by greater expression of smooth muscle actin (SMA; Extended Data Fig. 2a). Levels of vimentin, SMA and fibronectin expression further distinguished fibroblasts and myofibroblasts. Four fibroblast phenotypes expressed CD68 in the absence of CD45, in line with previous reports[15]. Comparable stromal diversity in breast cancer has recently been reported[16]. For epithelial phenotypes, key distinguishing

[1]Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland. [2]CRUK Cambridge Institute, University of Cambridge, Cambridge, UK. [3]Cambridge Breast Unit, Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge, UK. [4]A list of members and affiliations appears at the end of the paper. [5]Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, British Columbia, Canada. [6]These authors contributed equally: H. Raza Ali, Hartland Jackson. ✉e-mail: carlos.caldas@cruk.cam.ac.uk; bernd.bodenmiller@imls.uzh.ch

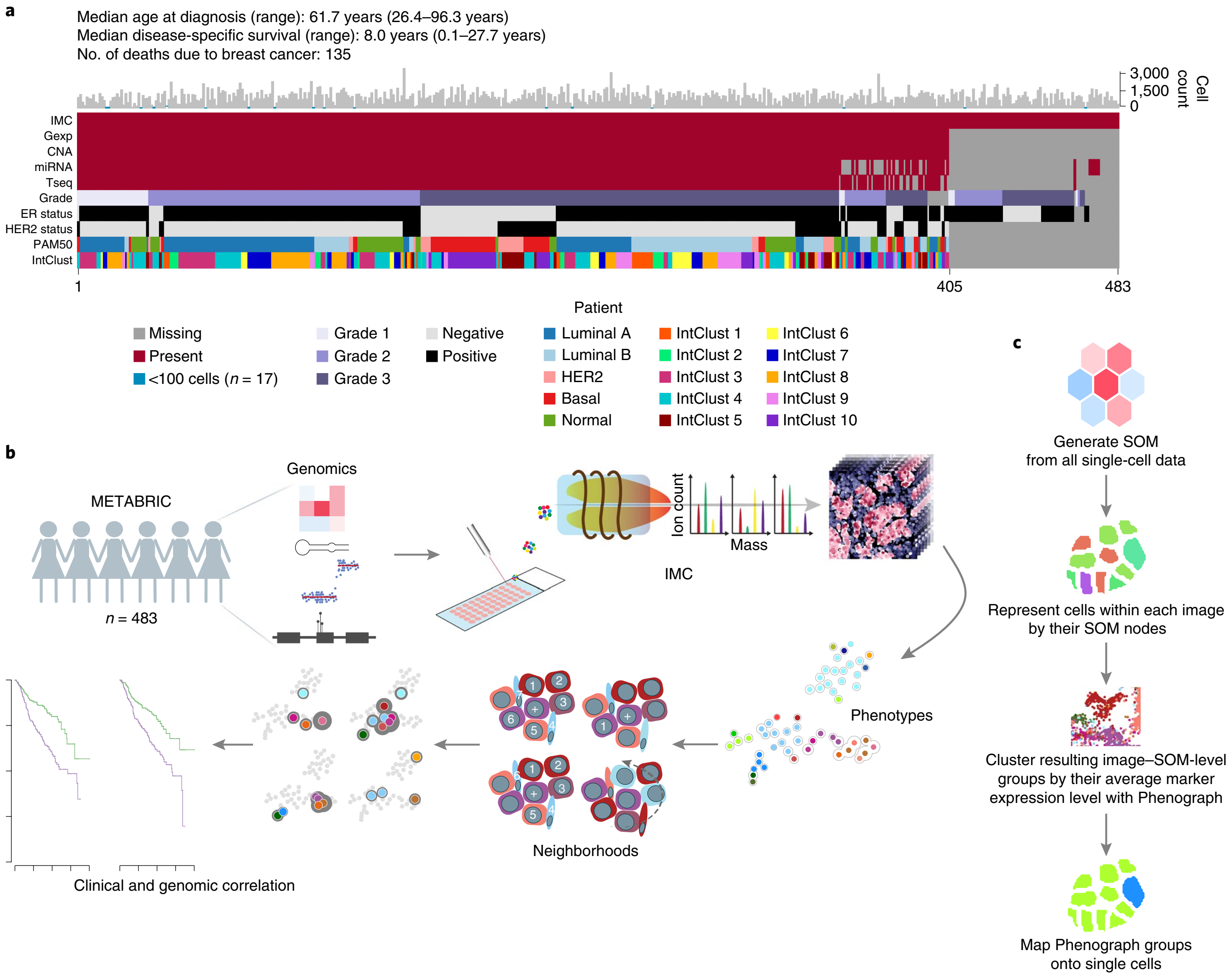

a
Median age at diagnosis (range): 61.7 years (26.4–96.3 years)
Median disease-specific survival (range): 8.0 years (0.1–27.7 years)
No. of deaths due to breast cancer: 135
3,000
1,500
0
Cell count
IMC
Gexp
CNA
miRNA
Tseq
Grade
ER status
HER2 status
PAM50
IntClust
1
405
483
Patient
Missing
Present
<100 cells (n = 17)
Grade 1
Grade 2
Grade 3
Negative
Positive
Luminal A
Luminal B
HER2
Basal
Normal
IntClust 1
IntClust 2
IntClust 3
IntClust 4
IntClust 5
IntClust 6
IntClust 7
IntClust 8
IntClust 9
IntClust 10
b
METABRIC
n = 483
Genomics
Ion count
Mass
IMC
Phenotypes
Neighborhoods
Clinical and genomic correlation
c
Generate SOM from all single-cell data
Represent cells within each image by their SOM nodes
Cluster resulting image–SOM-level groups by their average marker expression level with Phenograph
Map Phenograph groups onto single cells

Fig. 1 | Workflow to yield highly multiplexed molecular maps of METABRIC tumors by using IMC. a, Map of samples ordered by availability of data across platforms. The bar chart depicts the number of segmented cells per tumor. Samples comprising fewer than 100 cells (blue bars) were excluded from tumor-level analyses. Gexp, gene expression; Tseq, targeted sequencing. **b**, Experimental workflow for multiplexed IMC of 37 proteins in breast tumor tissues, with associated genomic annotation and clinical data. Tissue microarrays were labeled with isotope-tagged antibodies and subjected to IMC to quantify bound antibody abundance at 1-μm resolution. Resulting multidimensional images were processed, single cells were segmented and cellular neighborhoods were quantified. **c**, Schematic of the two-stage cell-clustering approach based on a self-organizing map (SOM) and Phenograph.

features included expression of hormone receptors (HRs); cytokeratin 5 (CK5), CK7 and CK19; human epidermal growth factor receptor 2 (HER2); and carbonic anhydrase IX, a marker of hypoxia. We also identified T cells, B cells, macrophages, endothelial cells, myoepithelial cells and vascular smooth muscle cells (Fig. 2d).

Transcriptomic correlations corroborate cellular identities. To test the validity of the assigned cell phenotypes, we assessed correlations between the proportions of cell phenotypes and bulk gene expression profiles in each tumor. The number of correlated genes varied substantially between cell phenotypes (Fig. 3a). We conducted comparative pathway analysis of the most positively correlated genes in each phenotype (Fig. 3b). This revealed three families of enriched pathways: (1) a group of related cell cycle pathways active in epithelial cells, (2) genes necessary for formation of the extracellular matrix and collagen deposition, enriched among myofibroblasts, and (3) a group of genes related to antigen presentation, interferon-γ signaling and interactions between lymphoid and non-lymphoid cells that were associated with all four T cell phenotypes and B cells. Thus, transcriptomic correlations with cell phenotypes corroborated the cellular identities we assigned on the basis of IMC data.

miRNAs are critical regulators of cell phenotypes within tumors[9,17]. In contrast to gene expression, which was balanced for positive and negative correlations for a given cell phenotype, there was a trend toward positive correlations between miRNA levels and a subset of four stromal phenotypes (vascular smooth muscle cells and three myofibroblast phenotypes; Fig. 3c). Pathway analysis of the genes targeted by the miRNAs correlated with these phenotypes revealed extracellular matrix terms, including extracellular matrix organization and collagen biosynthesis, among the top pathways (Extended Data Fig. 2b). These observations suggest that miRNA-mediated gene regulation is more important among stromal cells, including myofibroblasts, than in other cell phenotypes.

Genomic subtypes of breast cancer are characterized by diverse tumor ecosystems. We next compared cell phenotype distributions

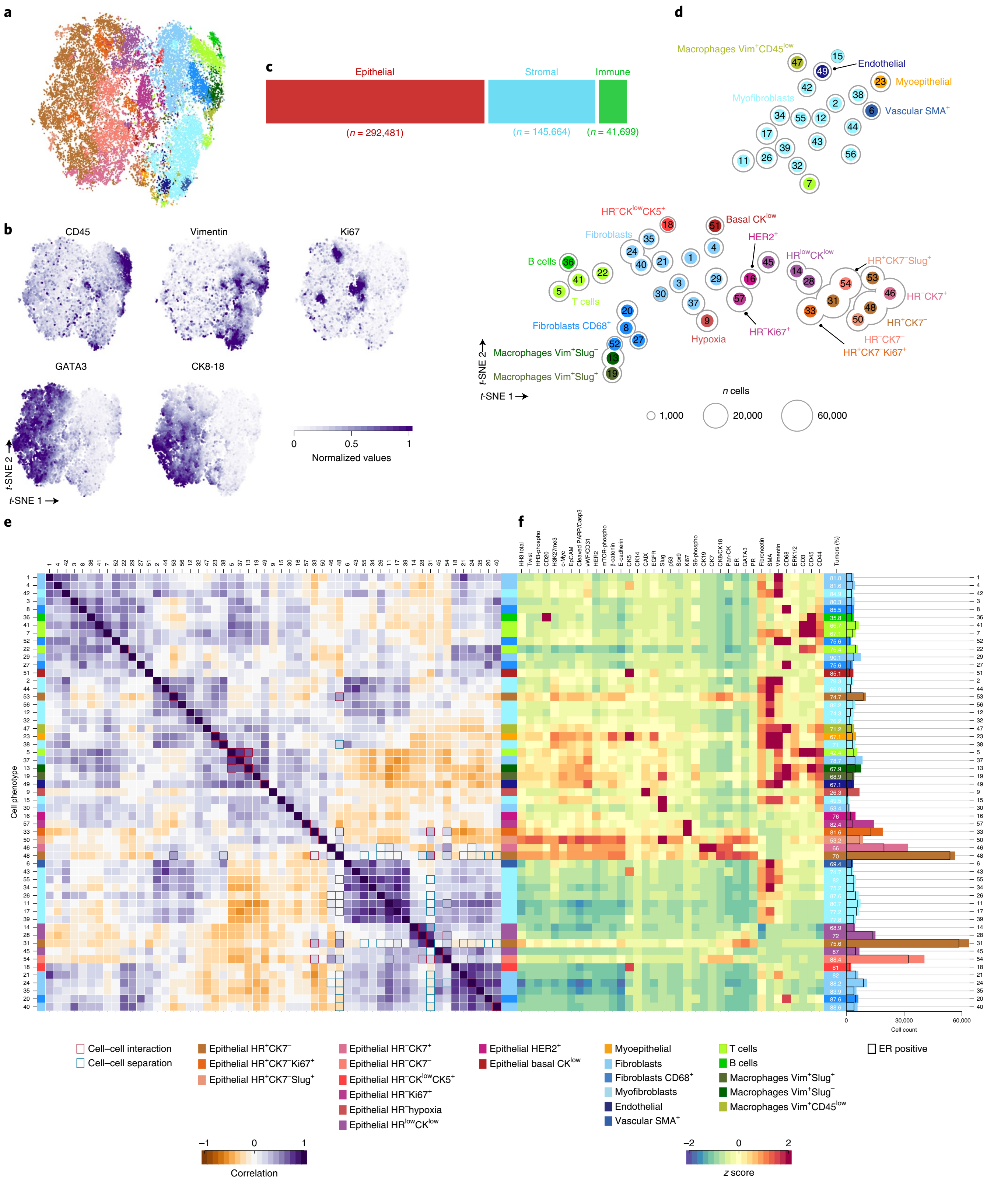


Fig. 2 | Data-driven derivation of cellular identities reveals composition of tumor ecosystems. a, Two-dimensional *t*-SNE representation of multiplexed proteomic data highlighted by cell phenotype. Each dot represents one cell; 5% of cells per tumor were randomly selected for illustration ($n = 24,003$ cells). **b**, *t*-SNE maps colored by expression of five canonical proteins. **c**, Bar plot showing the relative proportions of epithelial, stromal and immune cells for all cells analyzed. **d**, Annotated *t*-SNE map of cell phenotypes drawn using median protein expression levels. Vim, vimentin. **e**, Heat map of pairwise Spearman rank correlations of cell proportions, where the total cell count per tumor was taken as the denominator. Cell phenotypes were ordered by hierarchical clustering with Ward's method. Highlighted squares indicate significant cell–cell interactions (determined by permutation tests) observed in at least 10% of tumors. **f**, Heat map of median values of normalized protein expression per cell cluster. Markers were arranged by hierarchical clustering with Ward's method. The bar chart on the right depicts total cell count per cluster, distinguishing cells derived from ER$^+$ versus ER$^-$ tumors.

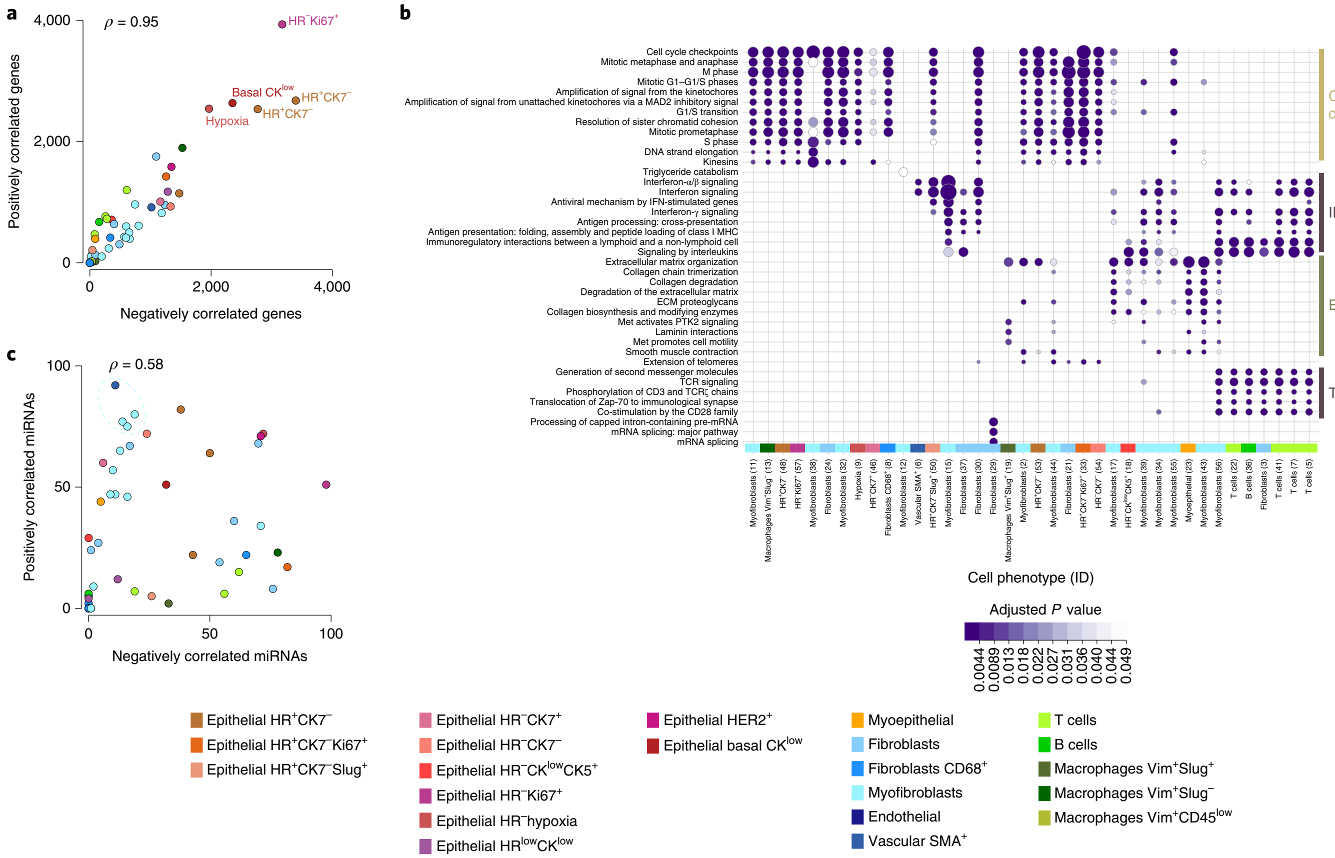


Fig. 3 | Transcriptomic correlations with IMC-defined cell types. a, Scatterplot comparing the number of positive and negative correlations between cell phenotype proportions and gene expression levels determined with linear models ($n = 390$ tumors). Each point represents one cell phenotype ($n = 55$ cell phenotypes; Spearman correlation = 0.95; $P < 0.05$). **b**, Comparative Reactome pathway enrichment analysis by cell phenotype based on the most strongly positively correlated genes ($n = 390$ tumors; hypergeometric test; *P* values are Benjamini–Hochberg adjusted for multiple comparisons). The top two terms per phenotype are depicted. Circle size is proportional to the number of genes associated with each term relative to the total number of genes per term. ECM, extracellular matrix; IFN, interferon; TCR, T cell antigen receptor. **c**, Scatterplot comparing the number of positive and negative correlations between cell cluster proportions and miRNA expression levels determined with linear models ($n = 371$ tumors). Each point represents one cell phenotype ($n = 55$ cell phenotypes; Spearman correlation = 0.58; $P < 0.05$).

and spatial features between breast cancer subtypes by linear regression. We focused on two widely used molecular taxonomies of breast cancer: the intrinsic molecular subtypes[1], based on tumor transcriptomes, and the integrative clusters[2], based on driver copy number aberrations (CNAs)[2].

We first investigated which of the cell phenotypes were enriched among different tumor subtypes. Several observations were consistent with prior knowledge, validating our approach. Epithelial cell phenotypes in particular showed distinctive enrichment patterns consistent with the known biology of the genomic subtypes (Fig. 4). Luminal A tumors were enriched for HR$^+$ epithelial cells (phenotypes 31, 48 and 53), whereas more proliferative luminal B tumors[1] were enriched for both HR$^+$ epithelial cells (phenotype 31) and HR$^+$Ki67$^+$ cells (phenotype 33). Basal-like tumors, which are mostly triple negative, showed enrichment of HR$^-$Ki67$^+$ cells (phenotype 57), epithelial cells expressing basal CKs (phenotype 51) and the phenotype associated with hypoxia (phenotype 9). Similarly, HR$^+$ cell phenotypes (31, 48 and 53) were enriched among the ER$^+$ integrative clusters (IntClusts 3, 4+, 6, 7 and 8), whereas IntClust 10 tumors, which map to the basal-like subtype, showed a cell enrichment pattern nearly identical to that of basal-like tumors. As expected, epithelial cells characterized by high expression of HER2 (phenotype 16) were enriched among the HER2 subtype and IntClust 5 tumors, defined by *ERBB2* amplification.

We also made several observations that highlight unexpected differences in the phenotypic composition of tumor subtypes. For instance, luminal subtypes were distinguished by their enrichment profiles for six key epithelial phenotypes (14, 28, 31, 46, 48 and 53) that varied in their expression of CKs and HRs (Fig. 4). Luminal B tumors were enriched for phenotypes 14 and 28, which had low HR and CK expression. IntClusts 2 and 6 also showed enrichment for cell phenotype 28. Cell phenotype 31, enriched in both luminal A and B tumors, also differed from phenotype 48 (only enriched in luminal A tumors) by lower expression of both HRs and CKs (Fig. 2f). This suggests that luminal B tumors have deviated further from a prototypical luminal epithelial cell than have luminal A tumors.

IntClusts 3, 4+, 7 and 8 were all characterized by enrichment for cell phenotype 48; all show low-to-intermediate genomic instability. IntClusts 7 and 8 have loss of 16q in common. IntClusts 6 and 8 were enriched for cell phenotype 31 despite their disparate genomic profiles (IntClust 6 tumors are characterized by the 8p12/*ZNF703* amplicon and IntClust 8 tumors by 1q gain/16q loss) and otherwise distinctive cell enrichment profiles. Cell phenotype 46, the only luminal cell phenotype to show high expression of both CK7 and CK19, was enriched among HER2 and IntClust 3 tumors. IntClust 3 tumors are characterized by few CNAs, frequent mutations of *PIK3CA*, *CDH1* and *RUNX1*, and the most favorable prognosis of all the IntClusts.

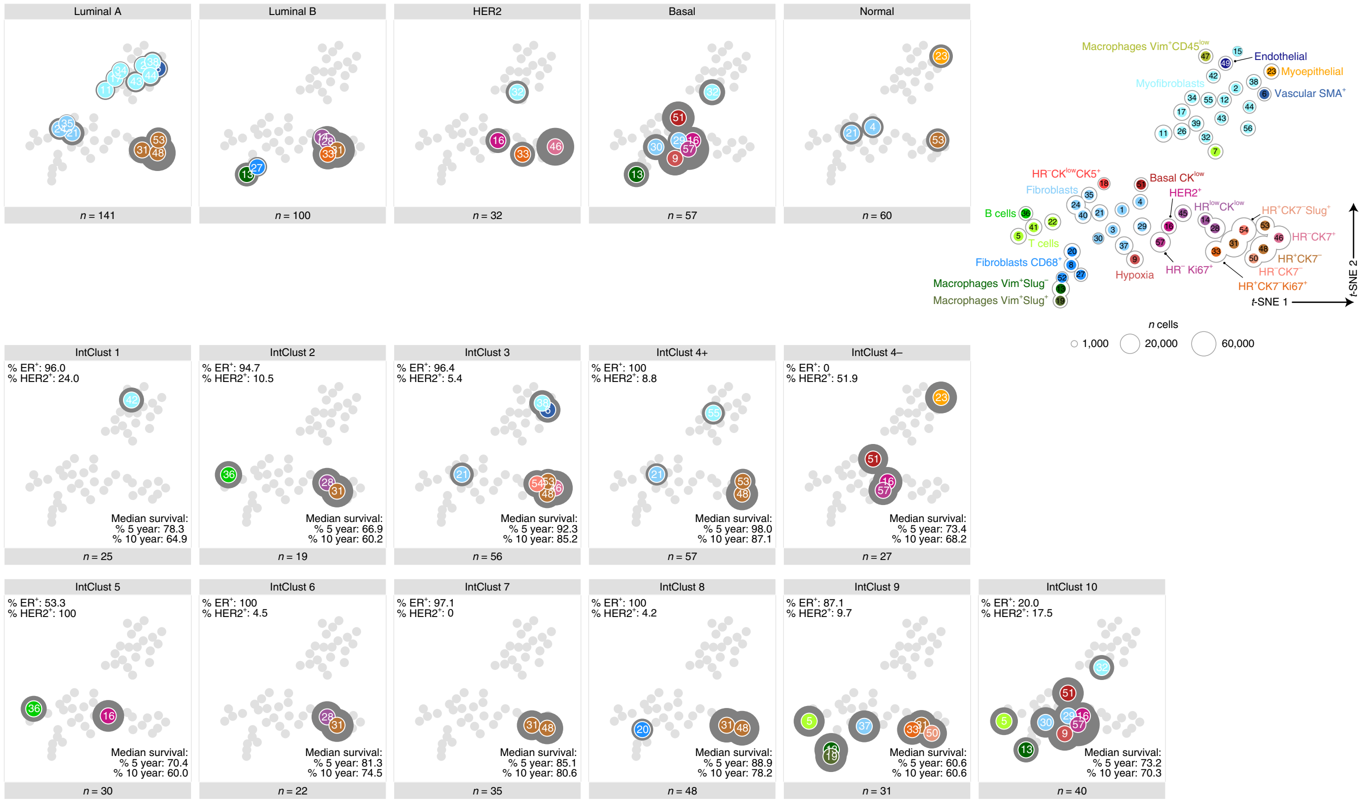


Fig. 4 | Phenotype enrichment in genomic breast cancer subtypes. Enriched phenotypes in each indicated genomic subtype are illustrated as two-dimensional *t*-SNE maps. The schematic map (right) indicates position by cell phenotype. Depicted associations were identified by linear regression, are limited to positive associations and are restricted to those associated at $P < 0.05$ (two-sided, adjusted for multiple comparisons per subtype by Benjamini–Hochberg correction). The dark gray background is proportional to the model coefficient, providing an indication of the strength of the association.

Enrichment for cell phenotype 46, which showed a highly distinctive expression profile, may indicate that the founding cell of these tumors occupies a different place in the mammary epithelial developmental lineage[18] than the founders of other IntClust tumor types.

We observed distinct patterns of stromal cell enrichment in different cancer subtypes (Fig. 4). Fibroblasts that expressed CD68 were enriched among estrogen receptor (ER)$^+$ luminal B tumors with poorer prognosis. Myofibroblasts were enriched in indolent ER$^+$ tumors (luminal A and IntClusts 3 and 4+), which are characterized by favorable prognosis but distinct genomic landscapes. Myofibroblasts were also enriched in IntClust 1 tumors, which are defined by the 17q23 amplicon. The enrichment patterns differed: IntClust 3 tumors, which harbor few CNAs but frequent mutations of *PIK3CA* and *CDH1*, were enriched for myofibroblast phenotype 38 and vascular smooth muscle cell phenotype 6. IntClust 4+ tumors, defined by few genomic aberrations, were enriched for cells of myofibroblast phenotype 55. Fibroblast phenotypes also showed distinct enrichment patterns among indolent ER$^+$ tumors: luminal A tumors were enriched for three fibroblast phenotypes (21, 24 and 35), whereas IntClust 3 and 4+ tumors shared enrichment of fibroblast phenotype 21. Myofibroblast cell phenotype 32 and fibroblast phenotypes 29 and 30 were enriched in both basal-like and IntClust 10 tumors, which often have *TP53* mutations. In summary, these findings indicate that genomically defined breast cancer subtypes contain distinct stromal cell repertoires.

We noted both T cell and macrophage enrichment among basal-like and IntClust 10 tumors. This may be related to the high mutational burden, genomic instability and frequent *TP53* mutations associated with IntClust 10 tumors[5]. Luminal B tumors of the IntClust 9 group were the only ER$^+$ subtype characterized by both macrophage and T cell enrichment. IntClust 9 tumors have intermediate-to-poor prognosis, are characterized by 8q amplification and have the highest proportion of *TP53* mutations among ER$^+$ tumors[2], which may be a factor in eliciting an immune response. We evaluated the robustness of our overall findings to potential biases or errors in the analytical methods and found that cell enrichment patterns among tumor subtypes were not adversely affected by signal bleed-through or choice of clustering method used to identify cell phenotypes (Extended Data Figs. 3–5a).

Genomic tumor subtypes were also characterized by different cell–cell interactions. We used permutation testing to identify interactions between the 57 cell phenotypes that occurred more or less frequently than expected by chance[19] and then investigated which of these were significantly enriched among tumor subtypes. We distinguished between interactions involving cells of the same cell phenotype (homotypic neighbors) and interactions involving cells of different phenotypes (heterotypic neighbors). Subtypes significantly enriched for interactions included HER2, basal-like and IntClust 10 (Extended Data Fig. 5b). In basal-like and IntClust 10 tumors, we observed abundant homotypic relationships among both epithelial and stromal cells. These tumors are, therefore, distinguished from other subtypes by a starker separation between compartments. We evaluated this further by comparing the average number of homotypic neighbors per cell phenotype and across molecular subtypes (Extended Data Fig. 6). Basal-like and IntClust 10 tumors were associated with more homotypic interactions, also suggestive of a 'separation phenotype'. Collectively, our findings reveal that breast cancer genomic subtypes have diverse cellular compositions, including marked differences in stromal phenotypes and in patterns of cellular interaction.

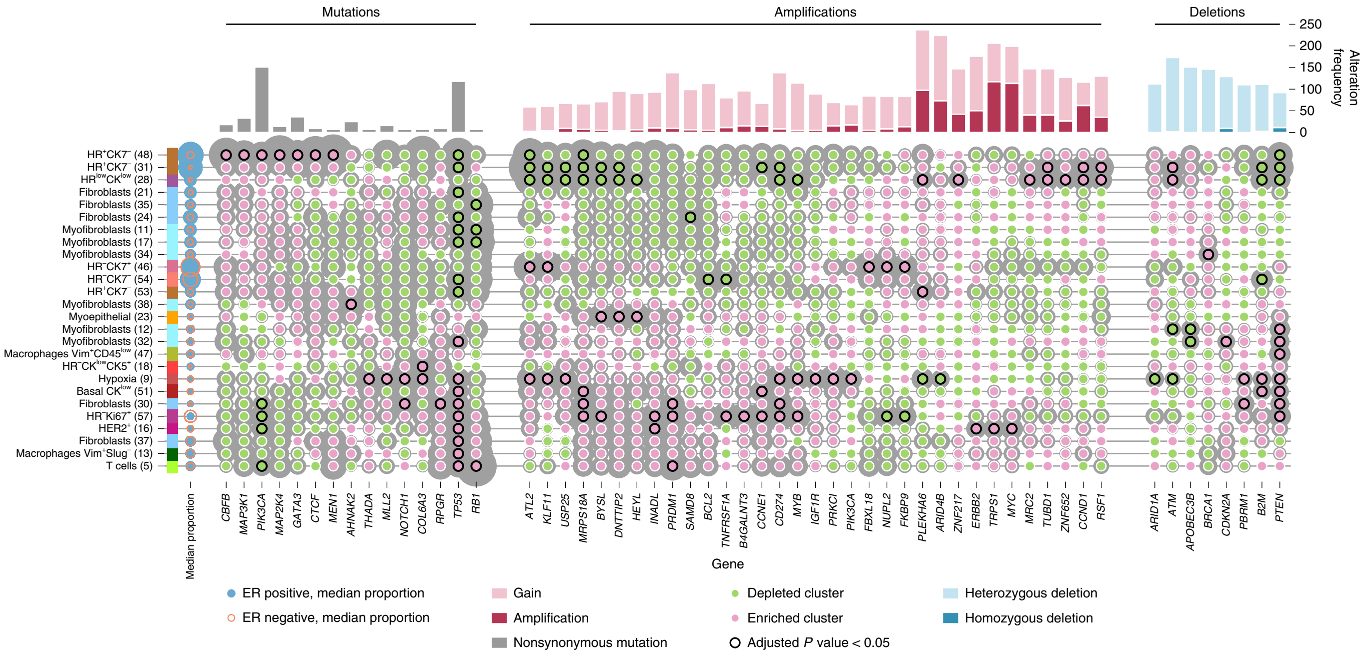


Fig. 5 | Somatic genomic alterations influence cell phenotypes. Patterns of association between cell phenotype proportions and driver somatic genomic alterations. Only phenotypes with at least one significant association at adjusted $P < 0.05$ are included. Associations were tested by linear regression ($n = 390$ tumors for CNAs and $n = 372$ tumors for mutations, two-sided tests). The gray background is proportional to the model coefficient, providing an indication of the strength of the association. Rows were ordered by hierarchical clustering of model coefficients; columns were ordered by hierarchical clustering of model coefficients within each aberration type (mutation, amplification or deletion). Bar charts depict the number of tumors with the corresponding alteration. Sizes of the leftmost markers labeled 'median proportion' are weighted by the median proportion of each cell phenotype by ER status.

Impact of somatic genomic alterations on breast tumor ecosystems. We next investigated associations between cell phenotype and somatic alterations in key driver genes[20]. We compared cell phenotype proportions between tumors with and without a particular alteration by linear regression (Fig. 5). We recovered relationships consistent with known breast cancer biology and also made several unexpected observations. For example, gains of *ERBB2* were associated with HER2$^+$ cell phenotype 16 (ref. [21]). Similarly, *TP53* mutation is known to occur more frequently among ER$^-$ tumors than in other types of breast cancer[5], and indeed we found that ER$^-$ basal cells (phenotype 51), hypoxia-associated epithelial cells (phenotype 9) and HR$^-$Ki67$^+$ epithelial cells (phenotype 57) were all positively associated with *TP53* mutations. In contrast, HR$^+$CK7$^-$ cells (phenotypes 31 and 48) were negatively associated with *TP53* mutations. For *PIK3CA*, the most frequently mutated oncogene in ER$^+$ breast cancer[5], this pattern was reversed: HR$^+$CK7$^-$ (phenotype 48) epithelial cells were positively associated with *PIK3CA* mutations, whereas HR$^-$Ki67$^+$ cells (phenotype 57) showed a negative association.

Cell phenotypes 28 (epithelial HR$^{\text{low}}$CK$^{\text{low}}$), 31 (epithelial HR$^+$; lower CK and HR expression) and 48 (epithelial HR$^+$CK7$^-$) were differentially enriched among luminal A and B tumors and were associated with distinct genomic events. Cell phenotype 48 was characterized by associations with more mutations than any other phenotype; these included mutations in *PIK3CA*, *GATA3*, *MAP3K1*, *CBFB*, *MAP2K4*, *CTCF* and *MEN1*. In contrast, cell phenotypes 28 and 31 were not associated with mutations, although these phenotypes were associated with CNAs including gains of *CCND1* and *TUBD1* and with *ATM* loss. These findings suggest that these ER$^+$ epithelial cell phenotypes are separated into those driven by mutations (phenotype 48) and those driven by CNAs (phenotypes 28 and 31).

The relationships that we uncovered in our analysis were not restricted to epithelial phenotypes. We found that fibroblast phenotypes 30 and 37 and myofibroblast phenotype 32 were associated with *TP53* mutations. Loss of *PTEN* was also associated with fibroblast phenotype 30 as well as myofibroblast phenotype 12. Other myofibroblast phenotypes showed negative associations with *TP53* and *RB1* mutations.

Next, we investigated associations between cell phenotypes and mutations in genes associated with immune cytolytic activity[22] to assess possible genomic selection for evasion of immune attack (Fig. 5). Epithelial cells that expressed carbonic anhydrase IX (phenotype 9), a marker of hypoxia, were associated with gains of *CD274*, which encodes PD-L1, and with heterozygous deletions of *B2M*, which encodes β2-microglobulin. This was the only cell phenotype positively associated with both of these alterations. This suggests that tumor cell hypoxia may enable selection of genomic alterations that facilitate immune evasion and supports the previously reported link between tumor hypoxia and an immune-tolerant microenvironment[23].

The genomic landscape of breast cancer is dominated by copy number events[4]; hence, we tested for associations between cell phenotype proportions and genome-wide CNAs (Extended Data Fig. 7). This analysis highlighted marked differences between cell phenotypes that would not be apparent without single-cell phenotypic data. For example, two luminal epithelial phenotypes, phenotypes 31 and 48, were both associated with gains of 16p. Phenotype 31, but not phenotype 48, was also correlated with loss of 11q. Despite the fact that both phenotypes 9 (hypoxia associated) and 57 (ER$^-$Ki67$^+$) were enriched among basal-like/IntClust 10 tumors, their CNA association profiles diverged substantially. Loss of 5q, a *trans* gene expression module specific to basal-like tumors that encodes key cell cycle and DNA repair genes[24], was a clear hallmark of phenotype 9, whereas gain of 10p, also characteristic of basal-like/IntClust 10 tumors, was a hallmark of phenotype 57.

We also assessed the relationship between cell phenotype abundance and genomic instability, calculated as the proportion of the genome affected by CNAs (Extended Data Fig. 8). This analysis showed that myofibroblast cell phenotypes 11 and 44 were inversely associated with genomic instability. In contrast, the proportions of $CD68^+$ fibroblasts (phenotype 8), proliferative epithelial cells (phenotypes 33 and 57), macrophages (phenotype 13) and T cells (phenotype 5) increased with genomic instability. Therefore, tumors with high genomic instability contain more proliferative cells and have distinctive stromal and immune populations.

To determine the overall contribution of different types of genomic information to cell phenotype composition, we investigated how much of the variance in cell phenotype proportions was explained by mutations, CNAs, and gene and miRNA expression. We addressed this by fitting a series of four linear models, each incremented by another data type (Extended Data Fig. 9). The explained variance for most cell phenotype proportions was substantially improved upon addition of gene expression data to mutation and CNA data but was not further improved upon addition of miRNA data. A set of stromal cells was an exception to this trend: for these cells, addition of miRNA data resulted in improvements in the explained variance for myofibroblasts (phenotypes 17, 34, 39 and 43), providing further support for the idea that miRNAs are more critical in regulation of gene expression in stromal cells than in other cell types in the tumor ecosystem. T cell abundance across all four T cell phenotypes was best explained by gene expression data, with little contribution from genomic alterations, in line with recent work[25].

Taken together, our systematic phenogenomic analysis indicates that somatic genomic aberrations exert influence over the cellular composition of both tumor cells and cells of the tumor microenvironment. We saw evidence for selective pressure of the immune response, and our data suggest that phenotypic features of tumor ecosystems, including hypoxia, are driven by a specific repertoire of large underlying genomic events that span genomic subtypes.

The prognostic impact of cell phenotypes depends on their genomic context. We examined whether the cell phenotypes and neighborhoods that we identified were predictive of clinical outcome and whether their prognostic effect differed among the IntClust subtypes. We conducted Cox regression analysis of cell phenotype proportions adjusted for ER status and plotted hazard ratios in rank order (Fig. 6a). To account for the compositional nature of the predictors (cell phenotype proportions), variables were modeled as log ratios taking myoepithelial and endothelial cells as referents for epithelial and non-epithelial cell phenotypes, respectively. As expected, cell phenotypes that expressed Ki67 (phenotypes 33 and 57) and HER2 (phenotype 16) were associated with poor outcome, as was the cell phenotype indicative of hypoxia (phenotype 9). Cells within the tumor microenvironment were also prognostic. Macrophages (phenotype 13) were indicative of poorer outcome, whereas vascular smooth muscle cells (phenotype 6) were associated with favorable prognosis. Phenotype 6 cells were enriched among luminal A and IntClust 3 tumors (Fig. 4).

To assess whether the spatial information in our dataset had prognostic relevance, we first investigated the correlations between cell phenotypes across all images (Fig. 6b). We annotated a correlation matrix of cell phenotypes with cell–cell interactions that occurred in at least 10% of images and used permutation testing[19] to distinguish whether cells were in contact more often (cell–cell interaction) or less often (cell–cell separation) with other cell phenotypes than expected by chance (Fig. 6b). The majority of interactions occurred between epithelial cells, either of the same or of different phenotypes. Cells of epithelial phenotypes 31 and 48 had negative interactions with fibroblasts and myofibroblasts. We observed patterns indicative of tumor microenvironment structure defined by both correlations (statistical sense) and interactions in an image (physical sense) between cell phenotypes[26].

Fibroblasts and myofibroblasts made distinctive contributions to tumor microenvironment structure. For example, one group on the heat map (Fig. 6b, square 1) showed correlations among T cells, macrophages and endothelial cells and an interaction between T cells and macrophages, but no stromal cells were involved in correlations or interactions. A stromal–lymphoid group, in contrast, involved correlations among fibroblasts, T cells and B cells and homotypic interactions among T cells (Fig. 6b, square 2). A third group composed of myofibroblasts and lacking an immune component involved both homotypic and heterotypic interactions (Fig. 6b, square 3). These patterns are suggestive of a spectrum of tumor microenvironments in breast cancer: at one end of the spectrum is a microenvironment characterized by diverse immune cells and endothelial cells; there is an intermediate microenvironment of lymphocytes and stromal cells; and, at the other end of the spectrum, there is an immune-depleted microenvironment dominated by myofibroblasts.

Next, we investigated the prognostic impact of cell neighborhoods, where a cell neighborhood was defined as the cells in contact with a given index cell. We used the mean of the number of homo- or heterotypic cell neighbors per cell phenotype per tumor, normalized to the number of neighboring cells, for survival analyses (Fig. 6c). Both homo- and heterotypic neighborhoods showed prognostic associations similar to those for the corresponding cell proportion predictor. An exception to this trend was the heterotypic neighborhood of myofibroblasts of phenotype 12 that was significantly associated with poor outcome; the proportion of this cell phenotype was not significantly associated with outcome. Finally, we evaluated the combined contributions of cell phenotypes and their neighborhoods to outcome prediction by fitting a multivariable Cox regression model by penalized maximum-likelihood estimation. Predictors selected by the model included homo- and heterotypic neighbors in addition to cell proportions (Fig. 6d,e and Supplementary Table 3), suggesting that spatial statistics such as neighborhoods may improve outcome prediction based on cell composition.

Finally, we investigated whether the prognostic effect of cell phenotypes significantly differed between IntClust subtypes (Fig. 6f). We identified three cell phenotypes, of which only one was of epithelial lineage (HR^+CK7^-, phenotype 48), that showed a significantly different prognostic effect within specific IntClust subtypes. Myofibroblasts of phenotype 55 were associated with favorable outcome among IntClust 1 tumors but not among other subtypes. These findings support a model of cancer-associated stroma as a constraint on tumor progression and suggest that this may be related to non-cell-autonomous effects of specific genomic alterations.

Cell phenotype 48, characterized by high CK and HR expression, was associated with favorable outcome among IntClust 6 tumors but not others. Notably, most IntClust 6 tumors, which are driven by 8p12 amplification, were enriched for tumor cell phenotypes with low CK and HR expression (phenotypes 28 and 31; Fig. 4). Phenotype 48 cells were characterized by associations with several mutations but not with CNAs, in contrast to phenotypes 28 and 31, which showed associations with CNAs but not with mutations (Fig. 5). Therefore, the subtype-specific prognostic effect of cell phenotype 48 may be related to intratumoral genetic heterogeneity among IntClust 6 tumors.

The only immune cell phenotype to demonstrate a subtype-specific prognostic effect was phenotype 13 ($vimentin^+Slug^-$ macrophages), which was associated with a favorable outcome among IntClust 7 tumors but with a poorer outcome among other subtypes. IntClust 7 tumors are characterized by 16p gain, 16q loss and mutations in *MAP3K1* and *CTCF* and were enriched for cell phenotype 48 (HR^+CK7^-). This supports previous observations of subtype-specific prognostic effects of immune cells such as macrophages in breast cancer[27,28]. These data show that IMC-derived cell phenotypes are linked to clinical outcome, illustrate the potential for identifying multiparametric tissue biomarkers by integrating multidimensional

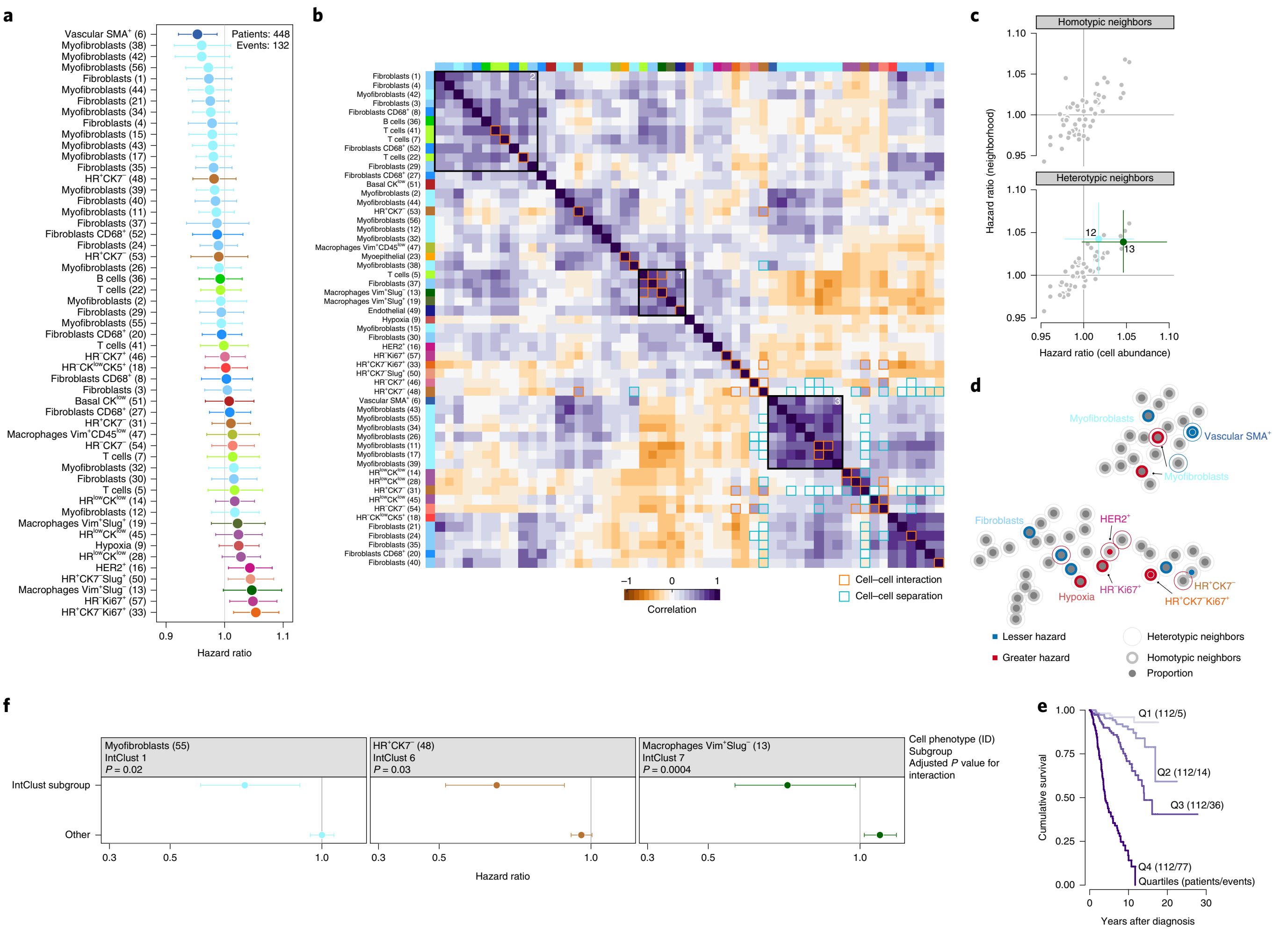


Fig. 6 | Prognostic impact of cell phenotypes and their neighborhoods. **a**, Hazard ratios of disease-specific survival for each cell type, modeled as log ratios. Circles represent point estimates, and whiskers indicate the 95% confidence interval derived from a Cox proportional-hazards model, adjusted for ER status ($n = 448$ patients, two-sided tests). **b**, Heat map of Spearman rank correlations among the proportions of each cell type, relative to all cells analyzed per tumor ($n = 467$ tumors). Cell phenotypes were ordered by hierarchical clustering. Highlighted orange and blue squares of cell–cell relationships are restricted to significant cell–cell interaction (orange) or separation (blue) as determined by permutation tests that were observed in at least 10% of images. **c**, Scatterplots comparing hazard ratios of cell neighborhoods and cell proportions (Cox regression, $n = 448$ patients). Those with a cell neighborhood $P < 0.05$ but a cell proportion $P > 0.05$ are highlighted. **d**, Phenotypes and neighborhoods selected by a multivariable model as predictors of disease-specific survival (regularized Cox regression, $n = 448$ patients). Colored markers represent features selected by the model, with red indicating an association with poorer outcome (greater hazard) and blue indicating an association with better outcome (lesser hazard); precise hazard ratios are provided in Supplementary Table 3. **e**, Survival plot by quartile of values (hazard ratios) predicted with the multivariable model depicted in **d**. **f**, Hazard ratio within one IntClust subgroup compared to the hazard ratio for all other IntClust subgroups combined for specified cell types. Depicted are those associated at $P < 0.05$ for interaction between cell phenotype and IntClust subtype (derived from a Cox regression model adjusted for ER status, $n = 390$ patients, P values adjusted for multiple comparisons).

single-cell data and quantitative spatial features, and reveal prognostic effects dependent on genomic context.

Discussion

We have conducted a phenogenomic analysis of cancer by integrating multidimensional breast tumor tissue imaging using IMC with multi-platform genomic data to investigate the impact of somatic alterations on tumor ecosystems at cellular spatial resolution. The tumor samples we studied were from the METABRIC cohort; these samples have been extensively characterized at the genomic level and are linked to long-term patient follow-up data[2,5,9,14]. We quantified the abundance of 37 epitope markers in each sample and used a data-driven approach to phenotypically classify cells and quantify cellular neighborhoods, revealing diverse tissue phenotypes that paralleled the genomic heterogeneity of breast cancer.

There was a separation of luminal epithelial cells into those associated with driver gene mutations but not CNAs (epithelial HR^+CK7^- cells of phenotype 48) and those associated with CNAs but few mutations (epithelial $HR^{low}CK^{low}$ cells of phenotype 28 and epithelial HR^+CK7^- cells of phenotype 31). Cells of phenotype 48 were enriched among ER^+ tumors with favorable prognosis (luminal A and IntClusts 3 and 4+) and were characterized by higher CK and HR expression than cells of phenotypes 28 and 31, which were enriched among luminal tumors of poor prognosis (luminal B and IntClust 6). Most luminal tumors were composed of a mixture of cell phenotypes rather than a single dominant population. This agrees with the observation that there is a continuum of proliferation rates among luminal tumors rather than a multimodal distribution[7,29]. Diverse transcriptional programs regulated by ER lead to the phenotypic diversity in luminal tumors[30,31]. Taken together with our findings, this suggests that the phenotypic compositions of luminal tumors are largely due to the interplay between somatic alterations and transcriptional programs induced by ER. Past work has suggested phenotypic expansion of minority cell populations under the pressure of endocrine treatment in luminal breast cancer[31]. This suggests that quantitative molecular mapping of cancer tissues, particularly longitudinal tracking of cell composition, may enable improved clinical decision-making.

IntClust 10 tumors, which are basal-like, had distinctive microenvironments defined by hypoxia and enrichment of T cells, macrophages and several stromal cell types. Of these, hypoxia-associated epithelial cells of phenotype 9 were associated with gains of *CD274* and loss of *B2M*, linking hypoxia to immune escape. Hypoxia has previously been linked to immune suppression[23,32]. The hypoxic environment may directly facilitate clonal diversity, possibly through impaired DNA-damage repair[33], or it may be a characteristic of tumors with high cell turnover and therefore more rapid clonal selection. As immune escape has been implicated in resistance to immune checkpoint blockade[34], markers of hypoxia may aid in identifying patients with de novo resistance or those likely to develop resistance to these agents.

Analysis of multidimensional tissue imaging data has challenges. Among them is how to accurately segment cells. Cancer tissues often contain areas of crowded cells such that it can be problematic to accurately separate one cell from another, and this may lead to mixing of signal between closely associated cells. We investigated the impact of different cell segmentation strategies by comparing whole-cell segmentation to a highly conservative annular approach that limited segmentation to a distance of up to three pixels from the nuclear edge. Cell phenotypic profiles were highly similar between these two approaches (Extended Data Fig. 3) but were not identical. Similarly, we compared different cell clustering strategies (Extended Data Fig. 4) and found largely concordant, but not identical, results. Our systematic investigation of these effects revealed that some variation in cell profiles and phenotypes can arise depending on which approach is adopted. Notably, the key findings were robust to these choices.

We uncovered unexpected diversity among stromal cells. Cancer-associated fibroblasts (CAFs) are typically described as expressing SMA, giving rise to the term myofibroblasts[35]. We observed these cells across tumors of all genomic subtypes, but they were most highly enriched in ER^+ tumors with low genomic instability. Survival analysis was suggestive of association of these cells with favorable outcome (Fig. 6a), in apparent disagreement with the putative protumoral role of CAFs. Myofibroblast phenotype 32 was an exception, as these cells were enriched in IntClust 10 tumors and were associated with high levels of genomic instability, more consistent with the prevailing CAF paradigm. There is, however, evidence to support our finding of CAF enrichment in tumors with favorable prognosis: The probable histopathological correlate of activated fibroblasts is stromal desmoplasia[35], a feature exemplified by pancreatic carcinomas, which are associated with a dismal prognosis, but for which CAFs have been implicated as cellular restraints of tumor progression[36,37]. In contrast, tubular carcinomas of the breast are also defined by marked stromal desmoplasia but have excellent prognosis[38]. A recent review of the METABRIC study revealed that tubular carcinomas belong to the IntClust 3 subtype[39], which was associated with enrichment of myofibroblast phenotype 38 in our analysis. Our findings therefore indicate that a subset of luminal breast tumors of favorable prognosis are characterized by fibroblast activation.

The cardinal features of the multicellular ecosystems of solid tumors have only begun to be explored. Here, by integrating multidimensional tissue imaging and multi-platform genomics data, we identified cellular phenotypic correlates of somatic genomic alterations and demonstrated their variable influence on tumor ecosystems. Our findings suggest that somatic genomic alterations collectively manifest as characteristic tumor ecosystems. Characterization of these ecosystems will further understanding of tumor evolution and will potentially enable identification of features that can be used to stratify patients and that can serve as targets for development of novel therapies.

Methods

Study population and genomic assays. We analyzed breast tumor samples from patients enrolled in the METABRIC study[2]. These patients were diagnosed with primary invasive carcinoma and treated in Cambridge, UK, between 1985 and 2005. Appropriate ethical approval from the institutional review board was obtained for the use of biospecimens with linked pseudo-anonymized clinical data. Extensive details of specimen handling, nucleic acid extraction, microarray hybridization, targeted sequencing and quality-control procedures have been described previously[2,5,9]. Briefly, nucleic acids were extracted from 30-µm sections from fresh frozen tissues with the DNeasy Blood and Tissue kit and the miRNeasy kit (Qiagen) on the QIAcube (Qiagen) according to the manufacturer's instructions. Genotyping and copy number analysis were conducted on Affymetrix SNP 6.0 arrays, and transcriptional profiling was conducted with the Illumina HT-12 v3 platform. Segmentation and copy number calls were made with circular binary segmentation, and gene expression data were normalized with the beadarray[40] R package. miRNA profiling was conducted on a custom Agilent microarray on which putative and known miRNA sequences were represented. For targeted sequencing, libraries were prepared with the Nextera Custom Target Enrichment kit (Illumina). Enrichment probes for 173 breast cancer driver genes were used to enrich for all exons. Samples were sequenced on an Illumina HiSeq 2000.

Tissue microarray construction and assessment of sampling error. Areas of invasive carcinoma suitable for in situ molecular analysis were identified on hematoxylin and eosin-stained slides by a breast pathologist (E.P.). Cores of 0.6 mm in diameter corresponding to marked areas were then removed and processed as previously described[41]. Of the 483 tumors included in this analysis, 463 were represented by one core, 19 were represented by two cores and 1 was represented by three cores. Where tumors were represented by more than one core, data from all cores were used to compute cell numbers and cell phenotype proportions. We used the subset of tumors represented by more than one core to assess whether sampling error was likely to prove problematic for our analysis (Extended Data Fig. 10). Our comparison was restricted to cores that contained at least 200 cells. We compared the cell phenotype composition of paired cores by hierarchical clustering. For 7 of the 15 samples with more than one core tested and containing at least 200 cells, the matched cores clustered together, indicating greater similarity between cores from the same tumor than between those from different tumors. Where cores from a tumor did not cluster together, this was often because the tissue content differed between them, for example, where one contained mostly stromal cells and the other contained mostly tumor cells. Therefore, although the study was not free of sampling error, these observations suggest that sampling error did not represent a major impediment.

Antibody conjugation. Descriptions of antibodies, isotope tags and concentrations used for staining are provided in Supplementary Table 2. Antibody–metal conjugation was conducted with the Maxpar labeling kit (Fluidigm). Following conjugation, the concentration was assessed with a NanoDrop (Thermo Scientific) and was adjusted to between 100 and 500 µg ml^{-1}. Antibodies were stored in Candor Antibody Stabilizer (Candor Bioscience) at 4 °C. The cloud-based platform AirLab was used for all antibody management and panel construction[42]. Antibody concentration and specificity were evaluated by visual inspection of IMC images of a variety of control tissues, including normal breast and invasive carcinoma.

Tissue antibody labeling. Slides were stained as previously described[19]. Briefly, slides were deparaffinized in xylene and rehydrated in a graded alcohol series. Antigen retrieval was conducted with Tris-EDTA (pH 9) buffer at 95 °C in a NxGen decloaking chamber (Biocare Medical). Following cooling, slides were blocked with 3% BSA in TBS for 1 h. Slides were then incubated with metal-tagged antibodies overnight at 4 °C with the exception of anti-ERα antibodies, which were detected with a metal-tagged anti-rabbit secondary antibody to increase signal (Supplementary Table 2). Following incubation, slides were washed with TBS. Finally, samples were incubated with 0.5 µM Cell-ID Intercalator-Ir (Fluidigm, 201192B) for detection of DNA. After 5 min, slides were rinsed with TBS and then air dried.

Imaging mass cytometry. The abundance of bound antibody was quantified with a Hyperion Imaging Mass Cytometer (Fluidigm). Tissue was laser ablated at 200 Hz. Ablated tissue aerosol was transported to a CyTOF mass cytometer (Fluidigm) for quantification, as previously described[8].

Image processing, single-cell signal quantification and identification of cell neighborhoods. Count data were converted to tiff image stacks and analyzed with a bespoke image processing pipeline (https://github.com/BodenmillerGroup/imctools). Briefly, random 125 × 125 µm^2 crops of images were generated and upscaled by a factor of 2 for pixel classification with the pixel-classification tool ilastik. Pixels were manually labeled as nuclear, cytoplasmic or background to train a random forest classifier in ilastik. The trained classifier was used to attribute probabilities to the remaining pixels, generating probability maps as RGB tiff files. We identified images where the pixel classifier was performing most poorly by quantifying the uncertainty of the classifier on each image; we then extended the training set with pixels from these images and repeated the process until improvement in model performance plateaued (four iterations). Probability maps were analyzed with CellProfiler[12]. Nuclei were detected as primary objects, and cytoplasm and cell membrane were identified by expanding primary objects to the border between the cell cytoplasm/membrane and background with the propagation method. Single-cell regions identified in this way formed a cell mask used for signal quantification and derivation of neighborhood relationships. Single-cell protein abundance estimates corresponded to the mean ion count of all pixels encompassed by a cell area. We adjusted for hot aggregates of antibody/metal in a manner similar to that previously described[43]. Briefly, we trained a pixel classifier to identify affected areas with ilastik, generated a corresponding mask and removed affected cells from analyses. We found that the majority of cells from the two rare phenotypes, phenotypes 10 and 25, were affected by hot pixel aggregates; hence, cells assigned to these were removed from analyses. We identified tissue showing 'edge effect' (a gradient of ion counts identifiable at the periphery of tissue spots) by manual inspection and isolated affected peripheral cells by using iterations of convex hulls to varying depth, as appropriate. Affected cells were removed from analyses.

Cell clustering. Single-cell expression data were arcsinh transformed with 0.8 as a cofactor before analysis. On the basis of protein distribution values across all cells, data were clipped at the 99th centile, and cells were included in clustering. Markers used for clustering were limited to the most informative in distinguishing cell populations and those deemed to have an acceptable signal-to-noise profile: CK8/CK18, CK19, CK5, CD68, CD3, CD20, ER, PR, CD45, GATA3, CK7, Ki67, SMA, HER2, pan-CK, EGFR, TP53, β-catenin, vWF/CD31, CAIX, Slug and vimentin. We analyzed data in two stages. First, we clustered cells into 225 groups with a self-organizing map[44] (15 × 15) implemented in the FlowSOM package[44], and then, by using the mean expression values within each of these clusters for each image, we conducted a second round of clustering with the community detection algorithm Phenograph[45], resulting in 57 clusters (of which 2 were removed following adjustment for hot pixel aggregates). These clusters were mapped back to single cells. To give these phenotypes descriptive labels, we used the average protein expression profile for each cluster to determine cell lineage on the basis of markers of epithelial (pan-CK, CK7, CK8/CK18, CK19), stromal (vimentin, fibronectin, SMA) and immune (CD45, CD3, CD20, CD68) cell types. Where average expression profiles were ambiguous with respect to these markers, images were also inspected to determine the most appropriate cell label on the basis of cell location and morphology.

Cell–cell interactions and cell neighborhoods. We used a previously described permutation testing approach[19] to determine whether interactions between cell phenotypes were observed more frequently than expected by chance. Briefly, the immediate neighbors of each cell as defined in the 'Object Relationships' table created with the CellProfiler pipeline were used to generate a null distribution of cell interaction frequencies by permuting cell labels 1,000 times for each image. The observed frequency of each interaction phenotype was compared to this null distribution. A *P* value was computed as the proportion of permuted frequencies with a value equal to or greater than the observed frequency, adding 1 to each side of the equation to avoid spurious *P* values of zero[46]. Whether a cell–cell relationship was deemed significant separation or interaction was determined by whether the observed frequency fell on the lower or upper tail of the null distribution, respectively. Adjustment for multiple testing was conducted for each image with the Benjamini–Hochberg method[47,48]. Cell neighborhood statistics were computed for each tumor as the average number of adjacent homo- or heterotypic neighbors per cell, adjusted for the number of neighbors. Homotypic neighborhood statistics were computed as the average number of cell neighbors that were of the same cell phenotype, and heterotypic neighborhood statistics were computed as the average number of cell neighbors that were of a different phenotype.

Statistics and reproducibility. Cell phenotypes were treated as proportions. Spearman rank correlations were computed on the basis of the proportion of a cell phenotype compared to all of the cells in a tumor. For comparison to genomic and clinical data, cell phenotype proportions were computed separately according to whether a cell was epithelial or not epithelial. Adjustments for multiple testing were conducted with the Benjamini–Hochberg method[47,48]. No statistical method was used to predetermine sample size. Samples comprising fewer than 100 cells were removed from tumor-level analyses. Cells affected by staining artifacts were removed from analyses. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Molecular subtypes. Intrinsic tumor subtypes were determined with the PAM50 method as previously described[2,49]. IntClust subtype was based on the original designation[2]. Enrichment of cell phenotypes by molecular subtype was tested separately for each subtype with a linear model. Logit-transformed cell type proportion (logit(proportion + 0.001)) was taken as the dependent variable with the subtype of interest represented by an indicator variable. Association between cell–cell interactions detected by permutation testing and molecular subtypes was conducted with logistic regression by taking a given cell–cell interaction as the dependent variable. Adjustment for multiple testing was conducted for each subtype.

Sensitivity analyses: cell segmentation. It was possible that stromal cell enrichment among tumor subtypes was related to signal bleed from tumor cells into adjacent stromal cells in closely packed areas, where cell segmentation can be problematic. We therefore examined the composition of all neighboring cells for each cell phenotype (Extended Data Fig. 7a,b). This showed that most neighboring cells tended to be of the same cell phenotype or the same cell lineage. Although this was the case for most stromal cell phenotypes, some showed a higher proportion of epithelial or immune cell neighbors than others (Extended Data Fig. 7a). When we compared the composition of neighboring cells separately for genomic subtypes of breast cancer, we did not find that those identified as enriched were neighbored by a greater proportion of epithelial cells than those that were not significantly enriched (Extended Data Fig. 7b). Stromal cell enrichment patterns among tumor subtypes were not, therefore, due to inappropriate attribution of tumor cell signal to adjacent stromal cells. We further tested for the influence of signal bleed by systematic comparison of two cell segmentation strategies. The first strategy was the propagation method, used for the analyses described in the main text, where cell perimeters depend on a combination of the distance to the nearest nucleus and changes in the gradient of probability generated by machine learning-based pixel classification. In the second strategy, a mask was drawn around each nucleus up to a maximum distance of three pixels, not including background, resulting in a shrunken mask for each cell. We then compared the expression profiles of stromal cell phenotypes based on whole-cell and three-pixel segmentation limited to cells that mapped unambiguously. A clustered heat map (Extended Data Fig. 7c) revealed that stromal phenotype expression profiles based on whole-cell segmentation clustered together with three-pixel counterparts with only one exception (phenotype 3), which was separated by phenotype 30 with a highly similar profile. This showed that molecular profiles were robust to the cell segmentation strategy and corroborated our conclusion that cell phenotypes were not adversely affected by signal bleed from adjacent cells.

Sensitivity analyses: cell clustering. To determine whether associations between cell phenotypes and tumor subtypes were robust to the cell clustering method used to identify cell phenotypes, we used FlowSOM rather than the clustering strategy with Phenograph to cluster cells into 100 groups. We then mapped these groups to each of the 55 cell phenotypes identified by our original method on the basis of the similarity of their expression profiles (Extended Data Fig. 8a). Finally, we tested for associations between these mapped cell phenotypes and tumor subtypes to compare patterns of association between the original cell phenotypes and their mapped counterparts. To account for random initialization in the clustering algorithm, this process was repeated 100 times. The distribution of mapped groups was reflected in the total cell count per phenotype and most phenotypes were successfully recovered (Extended Data Fig. 8b). There was excellent concordance for associations with genomic tumor subtypes between the original cell phenotypes and the newly mapped groups (Extended Data Fig. 8b). In sum, these findings showed that patterns of association with tumor subtypes were robust to choice of clustering strategy.

We also investigated whether combining all cells belonging to cell phenotypes with the same descriptive label (for example, fibroblasts) would lead to a meaningful loss of information. We combined cell phenotypes with the same descriptive label and tested for enrichment patterns among tumor subtypes (Extended Data Fig. 9). Major enrichment patterns were reproduced, including those of most epithelial, stromal and macrophage cell phenotypes; however, this simplification came at the cost of resolution. For example, differential enrichment between cell phenotypes 31 and 48 among luminal tumors, distinct stromal cell phenotype enrichment profiles and T cell enrichment patterns were lost when cells were combined into coarser groupings. This demonstrated the advantage of retaining all cell phenotypes in accurately mapping the complexity of the distinct tumor ecosystems of different tumor subtypes.

Correlation of cell phenotype with gene or miRNA expression. Gene expression and miRNA data, processed and normalized as previously described[2,9], were used for these analyses. Where more than one probe mapped to a gene, the probe with the greatest variance across the dataset was selected. Cell phenotype correlations with gene expression and miRNAs were estimated by linear regression following logit transformation of the cell proportions. Cell phenotype was used as the dependent variable and expression as the independent variable. Significant correlations were identified following adjustment for multiple testing. Enrichment analysis of Reactome pathways was conducted with the ReactomePA[50] and ClusterProfiler[51] packages. For gene expression, these analyses included up to the top 300 positively correlated genes per cluster. For pathway analysis of miRNAs enriched among myofibroblasts, probable gene targets were first identified. This was conducted with a data-driven approach. Genes were considered likely targets of miRNAs if more than 5% of expression variance was explained by the miRNA based on the results of a generalized additive model fit to the entire METABRIC cohort, as previously described[9]. Pathways enriched among the resulting targets were identified as for the gene expression analyses.

Correlation of cell phenotype with genomic variation. Data processing and normalization were conducted as previously described[2,5]. Genomic instability was computed as the proportion of the non-diploid genome on the basis of ASCAT integer copy number calls[52]. Kruskal–Wallis tests were used to test for association between cell phenotypes and quartiles of genomic instability. Associations between cell phenotypes and CNAs were tested separately for gains/amplifications and heterozygous/homozygous deletions where tumors were coded as either positive or negative for a given CNA. A similar strategy was used to analyze associations with mutations. Tumors were deemed either positive or negative for a given mutation encompassing all nonsynonymous mutations; genes with fewer than five mutations observed were excluded from association analyses on the basis of data contained in Supplementary Table 4 of ref. [5]. A given CNA or mutation was tested for association with a cell phenotype by using a linear model, taking the logit-transformed cell proportion as the dependent variable. Tests for association with CNAs were adjusted for the total number of amplified or deleted genes per tumor and the total number of copy number events per tumor. These features were represented by three rank-transformed covariates, as previously described[22]. Tests for association with mutations were adjusted for the total number of detected mutations, also represented as a rank-transformed covariate. Tests for association with copy number status were limited to genes previously identified as likely amplicon drivers[20], those associated with immune cytolytic activity[22] and those designated as 'large deletions' in breast cancer within the COSMIC database[53]. Analyses of these genes were limited to either increased or decreased copy number status as appropriate. Adjustment was made for multiple testing for each alteration type.

Explained variation of cell phenotypes by genomic data. The degree to which cell phenotype abundance was explained by each genomic data type was investigated with a linear model, taking logit-transformed cell phenotype proportion as the response variable. We fit a series of four models, each incremented by an additional data type (mutations, CNAs, gene expression and miRNA expression), represented by their first 20 principal components such that the full model contained 80 predictors. To account for the variable number of predictors, we used the adjusted R^2 statistic as an indicator of explained variance.

Survival analyses. Analyses were based on updated clinical data available in ref. [14]. To account for the compositional nature of the cell phenotype data, we took myoepithelial and endothelial cells as referents for epithelial and stromal cells, respectively, to compute log ratios that were then used as explanatory variables in Cox regression models[54]. Analyses were adjusted for ER status. To account for known violations of the proportional-hazards assumption by ER[55], it was modeled as a time-varying covariate: an additional term was included in the model that was allowed to vary with the logarithm of time. To determine whether prognostic effects significantly differed between IntClust subtypes, we extended these models to include an indicator variable for IntClust subtype and an interaction term between cell phenotype and IntClust subtype. *P* values for the interaction term were adjusted by Benjamini–Hochberg correction. Evaluation of all log ratios and neighborhoods (163 predictors) in a multivariate model was conducted with a penalized maximum-likelihood estimated Cox regression model implemented in the R package glmnet[56]. Lambda was selected by cross-validation. All analyses were conducted with Stata SE version 14.2 and R[57].

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

Data availability

IMC data, including cell masks and processed single-cell data, have been deposited to the Image Data Resource (https://idr.openmicroscopy.org/) under accession code idr0076 (see https://idr.openmicroscopy.org/about/download.html). Previously published METABRIC copy number, gene expression, miRNA and targeted sequencing data that were reanalyzed here are available under accession codes EGAS00000000083, EGAS00000000122 and EGAS00001001753 at the European Genome–Phenome Archive (http://www.ebi.ac.uk/ega/). Updated METABRIC clinical data analyzed here are available as part of the supplementary information in ref. [14]. All other data supporting the findings of this study are available from the corresponding authors upon reasonable request.

Code availability

In-house image preprocessing scripts are available at https://github.com/BodenmillerGroup/imctools. Other analysis code is available from the authors upon request.

Received: 6 November 2019; Accepted: 14 January 2020;
Published online: 17 February 2020

References

1. Perou, C. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
2. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
3. Ali, H. R. et al. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biol.* **15**, 431 (2014).
4. Ciriello, G. et al. Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45**, 1127–1133 (2013).
5. Pereira, B. et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 11479 (2016).
6. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
7. Wagner, J. et al. A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell* **177**, 1330–1345 (2019).
8. Giesen, C. et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11**, 417–422 (2014).
9. Dvinge, H. et al. The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature* **497**, 378–382 (2013).
10. Schulz, D. et al. Simultaneous multiplexed imaging of mRNA and proteins with subcellular resolution in breast cancer tissue samples by mass cytometry. *Cell Syst.* **6**, 25–36 (2018).
11. Damond, N. et al. A map of human type 1 diabetes progression by imaging mass cytometry. *Cell Metab.* **29**, 755–768 (2019).
12. Carpenter, A. E. et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
13. Haubold, C. et al. Segmenting and tracking multiple dividing targets using ilastik. *Adv. Anat. Embryol. Cell Biol.* **219**, 199–229 (2016).
14. Rueda, O. M. et al. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature* **567**, 399–404 (2019).
15. Gottfried, E. et al. Expression of CD68 in non-myeloid cell types. *Scand. J. Immunol.* **67**, 453–463 (2008).
16. Costa, A. et al. Fibroblast heterogeneity and immunosuppressive environment in human breast cancer. *Cancer Cell* **33**, 463–479 (2018).
17. Mitra, A. K. et al. MicroRNAs reprogram normal fibroblasts into cancer-associated fibroblasts in ovarian cancer. *Cancer Discov.* **2**, 1100–1108 (2012).
18. Stingl, J. & Caldas, C. Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. *Nat. Rev. Cancer* **7**, 791 (2007).
19. Schapiro, D. et al. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nat. Methods* **14**, 873–876 (2017).
20. Akavia, U. D. et al. An integrated approach to uncover drivers of cancer. *Cell* **143**, 1005–1017 (2010).
21. Slamon, D. J. et al. Studies of the *HER-2/neu* proto-oncogene in human breast and ovarian cancer. *Science* **244**, 707–712 (1989).
22. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61 (2015).
23. Facciabene, A. et al. Tumour hypoxia promotes tolerance and angiogenesis via CCL28 and T_{reg} cells. *Nature* **475**, 226–230 (2011).
24. Dawson, S. J., Rueda, O. M., Aparicio, S. & Caldas, C. A new genome-driven integrated classification of breast cancer and its implications. *EMBO J.* **32**, 617–628 (2013).
25. Cristescu, R. et al. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. *Science* **362**, eaar3593 (2018).
26. Bodenmiller, B. Multiplexed epitope-based tissue imaging for discovery and healthcare applications. *Cell Syst.* **2**, 225–238 (2016).
27. Ali, H. R. et al. Association between $CD8^+$ T-cell infiltration and breast cancer survival in 12 439 patients. *Ann. Oncol.* **25**, 1536–1543 (2014).
28. Ali, H. R., Chlon, L., Pharoah, P. D., Markowetz, F. & Caldas, C. Patterns of immune infiltration in breast cancer and their clinical implications: a gene-expression-based retrospective study. *PLoS Med.* **13**, e1002194 (2016).
29. Reis-Filho, J. S. & Pusztai, L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet* **378**, 1812–1823 (2011).
30. Ross-Innes, C. S. et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
31. Patten, D. K. et al. Enhancer mapping uncovers phenotypic heterogeneity and evolution in patients with luminal breast cancer. *Nat. Med.* **24**, 1469–1480 (2018).
32. Barsoum, I. B., Koti, M., Siemens, D. R. & Graham, C. H. Mechanisms of hypoxia-mediated immune escape in cancer. *Cancer Res.* **74**, 7185–7190 (2014).
33. Bristow, R. G. & Hill, R. P. Hypoxia, DNA repair and genetic instability. *Nat. Rev. Cancer* **8**, 180–192 (2008).
34. Sade-Feldman, M. et al. Resistance to checkpoint blockade therapy through inactivation of antigen presentation. *Nat. Commun.* **8**, 1136 (2017).
35. Kalluri, R. & Zeisberg, M. Fibroblasts in cancer. *Nat. Rev. Cancer* **6**, 392–401 (2006).
36. Ozdemir, B. C. et al. Depletion of carcinoma-associated fibroblasts and fibrosis induces immunosuppression and accelerates pancreas cancer with reduced survival. *Cancer Cell* **25**, 719–734 (2014).

37. Rhim, A. D. et al. Stromal elements act to restrain, rather than support, pancreatic ductal adenocarcinoma. *Cancer Cell* **25**, 735–747 (2014).
38. Rakha, E. A. et al. Tubular carcinoma of the breast: further evidence to support its excellent prognosis. *J. Clin. Oncol.* **28**, 99–104 (2010).
39. Mukherjee, A. et al. Associations between genomic stratification of breast cancer and centrally reviewed tumour pathology in the METABRIC cohort. *NPJ Breast Cancer* **4**, 5 (2018).
40. Dunning, M. J., Smith, M. L., Ritchie, M. E. & Tavare, S. beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics* **23**, 2183–2184 (2007).
41. Kononen, J. et al. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat. Med.* **4**, 844–847 (1998).
42. Catena, R., Ozcan, A., Jacobs, A., Chevrier, S. & Bodenmiller, B. AirLab: a cloud-based platform to manage and share antibody-based single-cell research. *Genome Biol.* **17**, 142 (2016).
43. Keren, L. et al. A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell* **174**, 1373–1387 (2018).
44. Van Gassen, S. et al. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* **87**, 636–645 (2015).
45. Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
46. Phipson, B. & Smyth, G. K. Permutation *P*-values should never be zero: calculating exact *P*-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.* **9**, Article39 (2010).
47. Newson, R. B. Frequentist *q*-values for multiple-test procedures. *Stata J.* **10**, 568–584 (2010).
48. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
49. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
50. Yu, G. & He, Q. Y. ReactomePA: an R/Bioconductor package for Reactome pathway analysis and visualization. *Mol. Biosyst.***12**, 477–479 (2016).
51. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* **16**, 284–287 (2012).
52. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
53. Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
54. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* **8**, 2224 (2017).
55. Blows, F. et al. Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLoS Med.* **7**, e1000279 (2010).
56. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** (1), 1–22 (2010).
57. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2016).

Acknowledgements

H.R.A. was supported by a Cancer Research UK (CRUK) Clinician Scientist Fellowship. We thank N. De Souza for critical review of the manuscript. The research of B.B. is funded by an SNSF R'Equip grant, an SNSF Assistant Professorship grant, an NIH grant (UC4 DK108132) and by the European Research Council (ERC) under the European Union's Seventh Framework Program (FP/2007-2013)/ERC grant agreement 336921. The laboratories of B.B. and C.C. were supported by the CRUK IMAXT Grand Challenge for this work.

Author contributions

H.R.A., C.C. and B.B. conceived the study; IMC experiments were performed by H.W.J.; H.R.A. conducted data analysis with assistance from J.F.; H.R.A., C.C. and B.B. wrote the manuscript with contributions from E.D.; V.R.T.Z. built the preprocessing pipeline; H.B. and E.P. constructed the tissue microarrays; S.-F.C. and O.M.R. generated and processed genomic data; O.M.R. and E.P. collected and curated clinical data; C.C. and S.A. co-led the METABRIC consortium; C.C. and B.B. co-directed the study; and all authors read and approved the final manuscript.

Competing interests

C.C. is a member of the External Science Panel of AstraZeneca, and his laboratory has received research grants (administered by the University of Cambridge) from Genentech, Roche, AstraZeneca and Servier. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43018-020-0026-6.

Supplementary information is available for this paper at https://doi.org/10.1038/s43018-020-0026-6.

Correspondence and requests for materials should be addressed to C.C. or B.B.

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

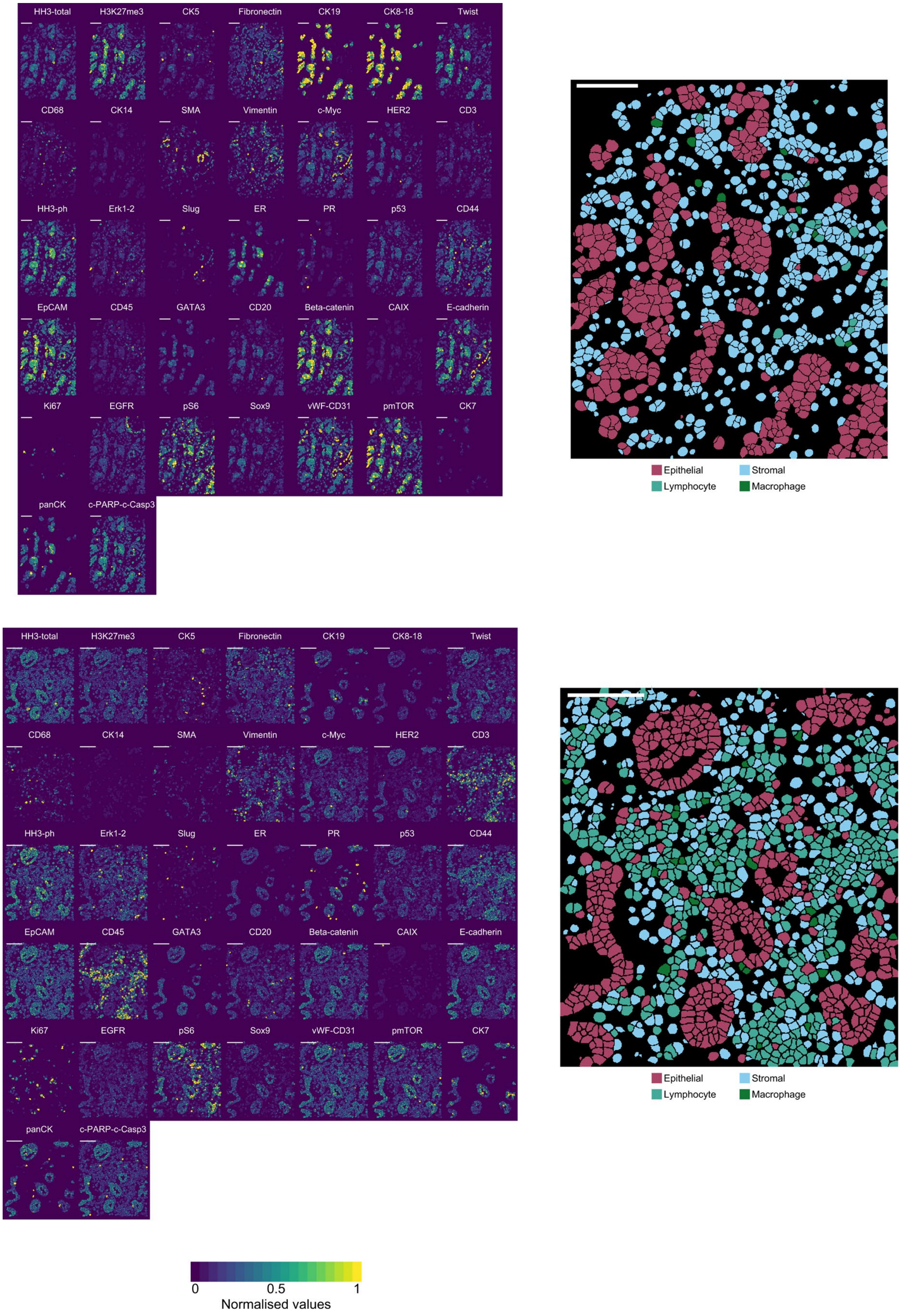
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

CRUK IMAXT Grand Challenge Team

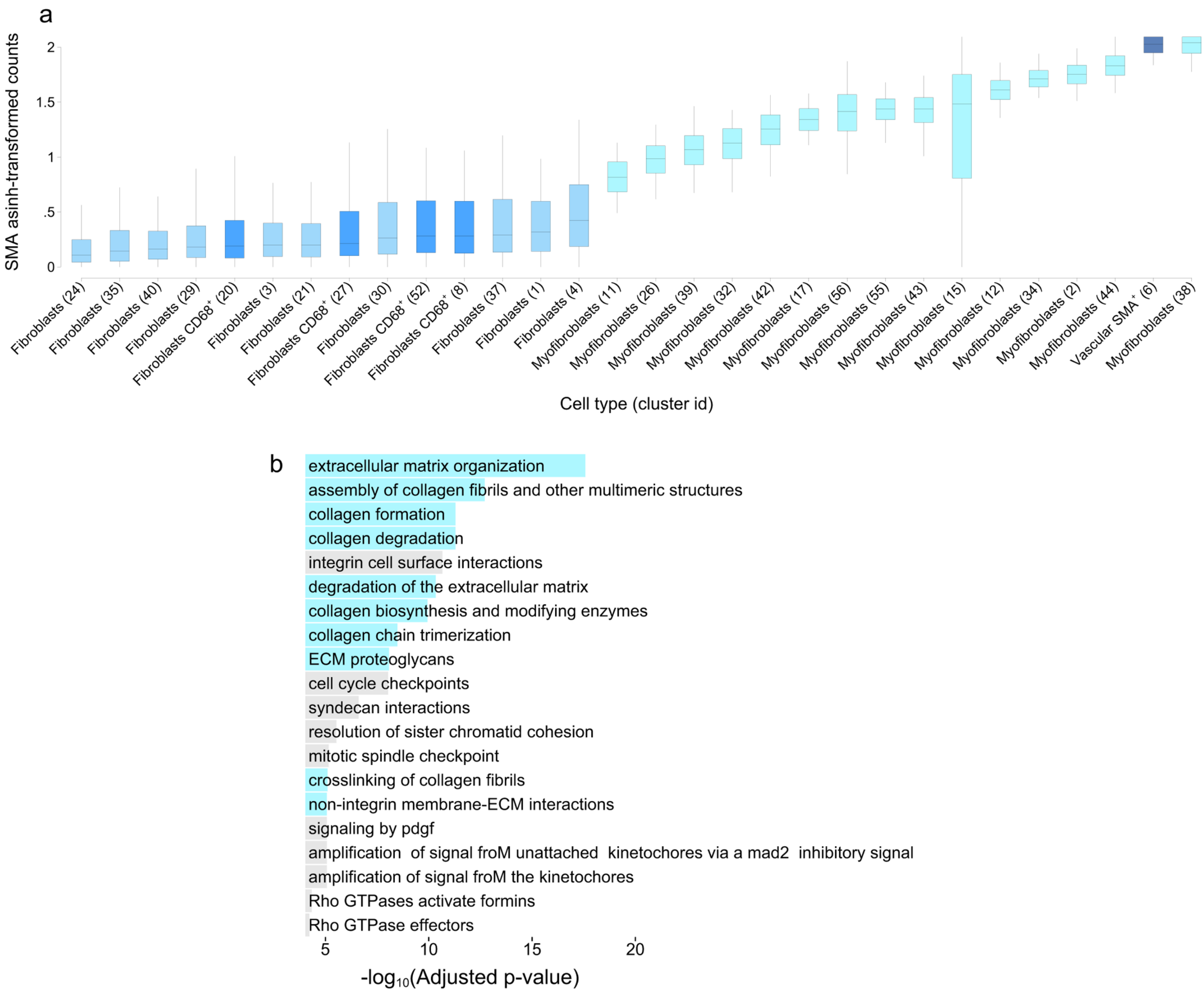
H. Raza Ali[7,8], M. Al Sa'd[9], S. Alon[10], Samuel Aparicio[11,12], G. Battistoni[7], S. Balasubramanian[7,13], R. Becker[14], Bernd Bodenmiller[8], E. S. Boyden[10], D. Bressan[7], A. Bruna[15], B. Marcel[8], Carlos Caldas[15], M. Callari[7], I. G. Cannell[7], H. Casbolt[7], N. Chornay[9], Y. Cui[10], A. Dariush[9], K. Dinh[16], A. Emenari[10], Y. Eyal-Lubling[15], J. Fan[17], E. Fisher[7], E. A. González-Solares[9], C. González-Fernández[9], D. Goodwin[10], W. Greenwood[7], F. Grimaldi[14], G. J. Hannon[7], O. Harris[14], S. Harris[14], C. Jauset[7], J. A. Joyce[18], E. D. Karagiannis[10], T. Kovačević[7], L. Kuett[8], R. Kunes[16], A. Küpcü Yoldaş[9], D. Lai[11,12], E. Laks[11,12], H. Lee[17], M. Lee[7,13], G. Lerda[7], Y. Li[11], A. McPherson[11,12,19], N. Millar[9], C. M. Mulvey[7], F. Nugent[7], C. H. O'Flanagan[11], M. Paez-Ribes[7], I. Pearsall[7], F. Qosaj[7], A. J. Roth[11,12,20], Oscar M. Rueda[15], T. Ruiz[11], K. Sawicka[7], L. A. Sepúlveda[17], S. P. Shah[11,12,19], A. Shea[15], A. Sinha[10], A. Smith[11], S. Tavaré[7,16,21], S. Tietscher[8], I. Vázquez-García[19], S. L. Vogl[14], N. A. Walton[9], A. T. Wassie[10], S. S. Watson[18], S. A. Wild[7], E. Williams[7], J. Windhager[8], C. Xia[17], P. Zheng[17] and X. Zhuang[17]

[7]CRUK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge, UK. [8]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [9]Institute of Astronomy, University of Cambridge, Cambridge, UK. [10]McGovern Institute, Departments of Biological Engineering and

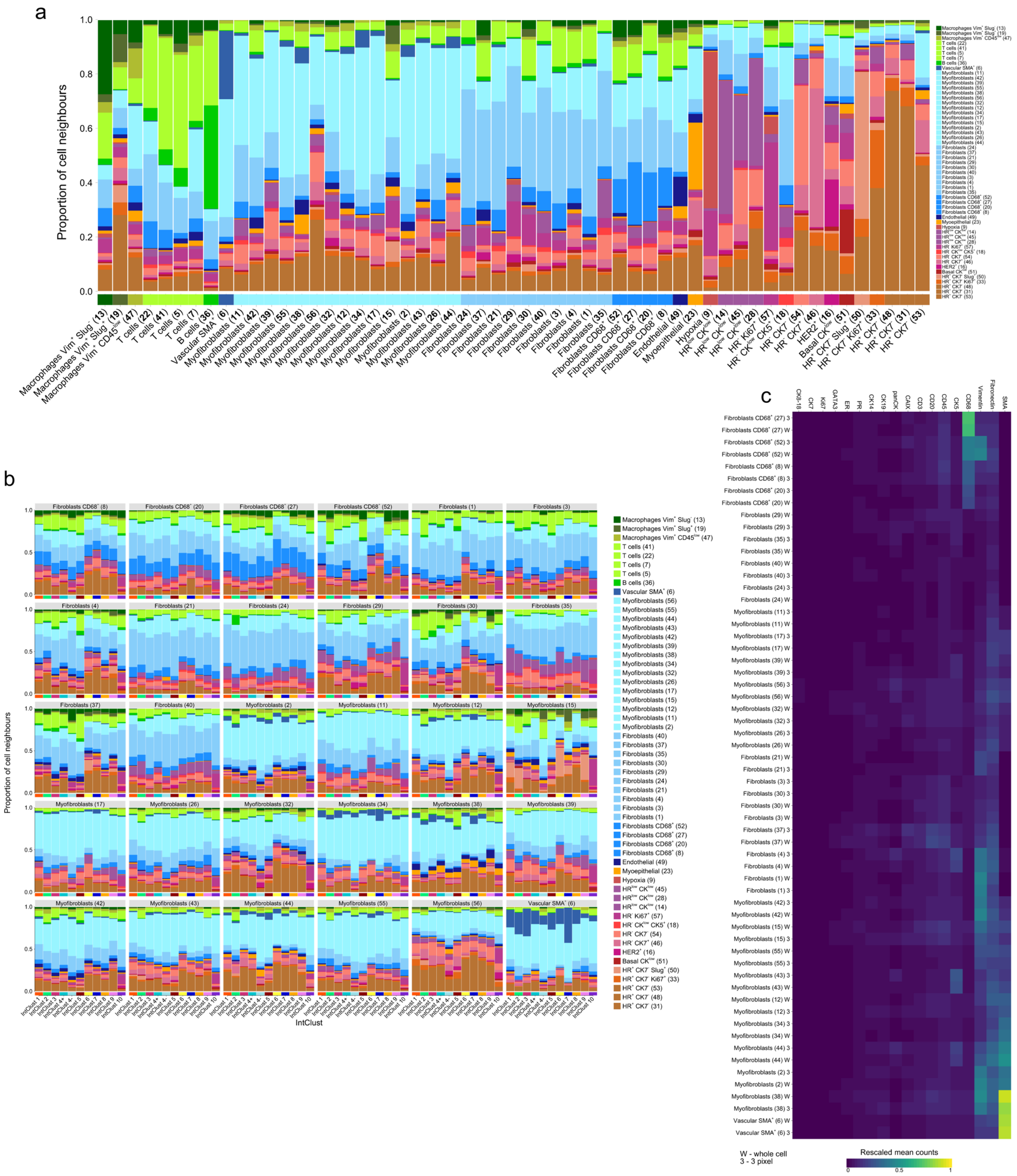
of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [11]Department of Molecular Oncology, BC Cancer, part of the Provincial Health Services Authority, Vancouver, British Columbia, Canada. [12]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada. [13]Department of Chemistry, University of Cambridge, Cambridge, UK. [14]Súil Interactive, Dublin, UK. [15]Department of Oncology and CRUK Cambridge Institute, University of Cambridge, Cambridge, UK. [16]Herbert and Florence Irving Institute for Cancer Dynamics, Columbia University, New York, NY, USA. [17]Howard Hughes Medical Institute, Department of Physics and of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA. [18]Department of Oncology and Ludwig Institute for Cancer Research, University of Lausanne, Lausanne, Switzerland. [19]Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [20]Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada. [21]New York Genome Center, New York, NY, USA.



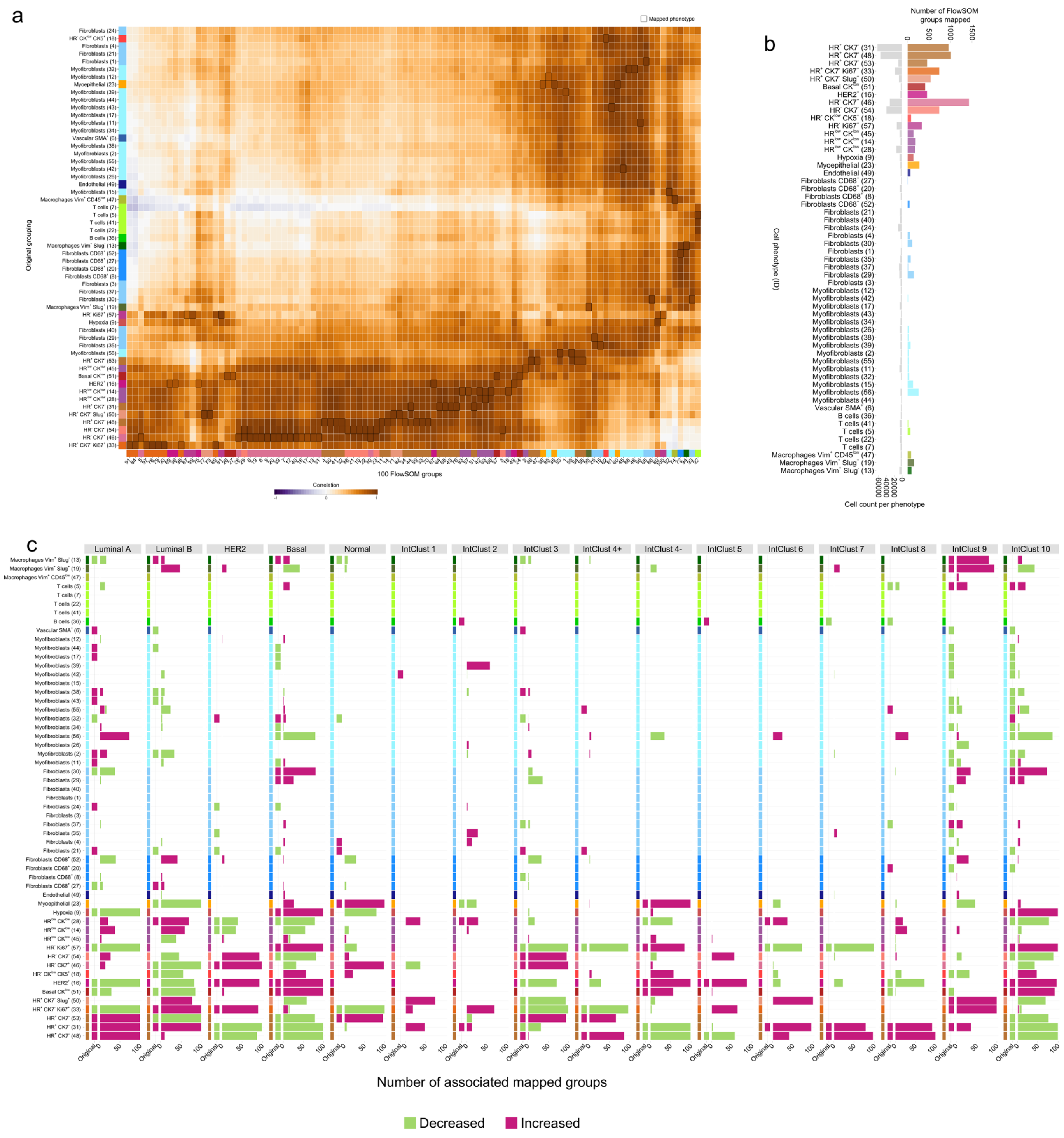
Extended Data Fig. 1 | Spatial distribution of cell types. Representative cell masks annotated by protein expression levels (left) and by inferred cell identities separated by tumour versus stroma (right). For protein expression levels, mean counts per cell were normalised relative to all cells analysed. White scale bars (top left) represent 100 µm.



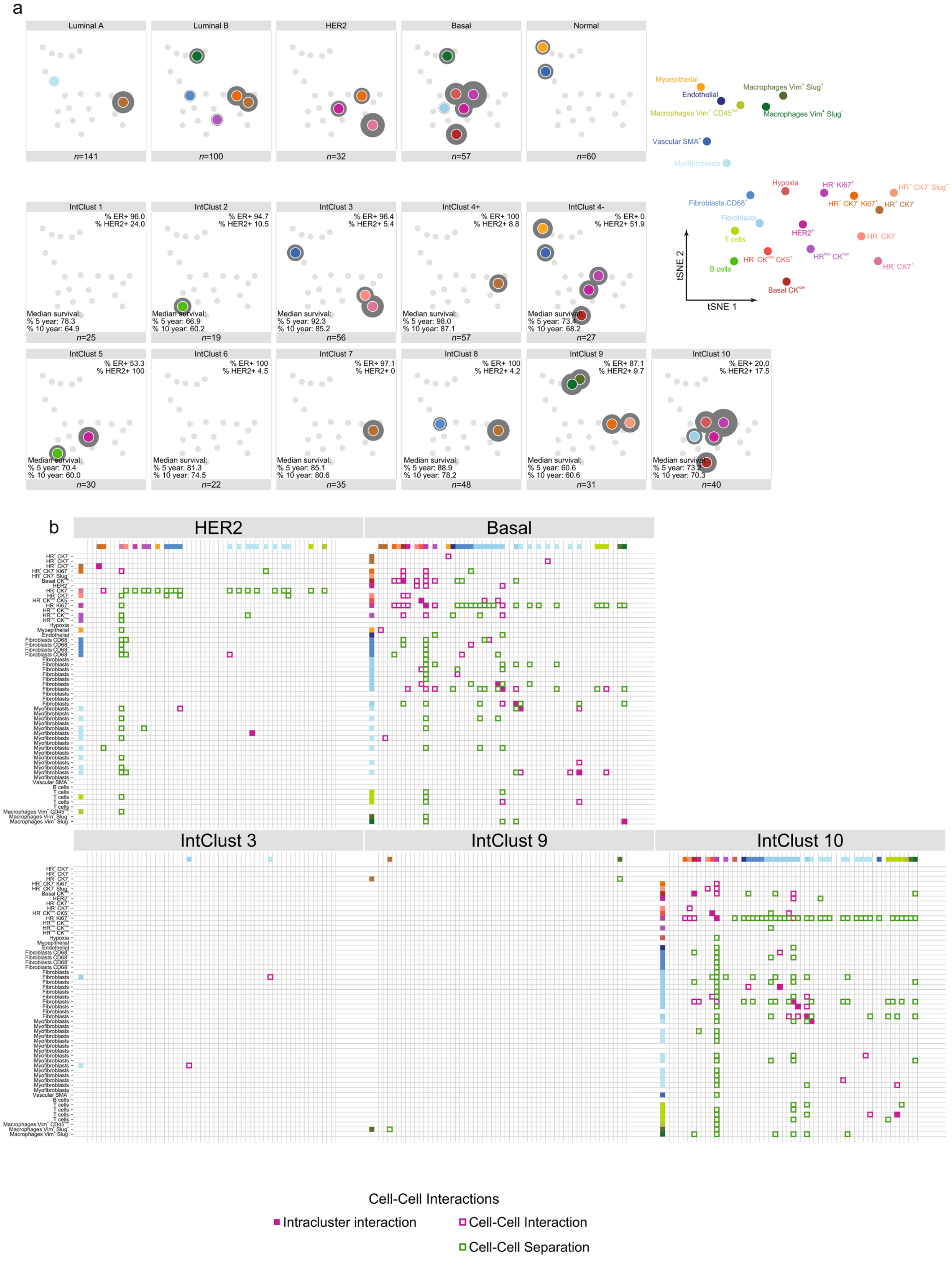
Extended Data Fig. 2 | Definition of myofibroblasts and miRNA-related pathway enrichment. a, Box-and-whisker plot of the distribution of SMA expression by stromal cell types (*n* = 141, 818 cells). Boxes represent the interquartile range. Lines dividing boxes indicate the median, and vertical lines represent range of expression from the 1st to 99th percentile. **b**, Pathway enrichment analysis of genes (hypergeometric test; adjusted for multiple comparisons) linked to miRNAs positively correlated with myofibroblast and vascular smooth muscle cell proportions (*n* = 371 tumours; related to Fig. 4).



Extended Data Fig. 3 | Cell neighbours depend on cell phenotype. a, Stacked bar plot depicting phenotypic composition of cell neighbours separately by each cell phenotype for all tumours. **b**, Stacked bar plots depicting the phenotypic composition of cells neighbouring all stromal phenotypes separately by IntClust subtype, illustrating patterns of stromal cell enrichment among IntClust subtypes (related to Fig. 4). **c**, Heatmap of median expression values for stromal cells based on both whole-cell and 3-pixel annular segmentation methods; rows and columns ordered by hierarchical clustering using Ward's method.

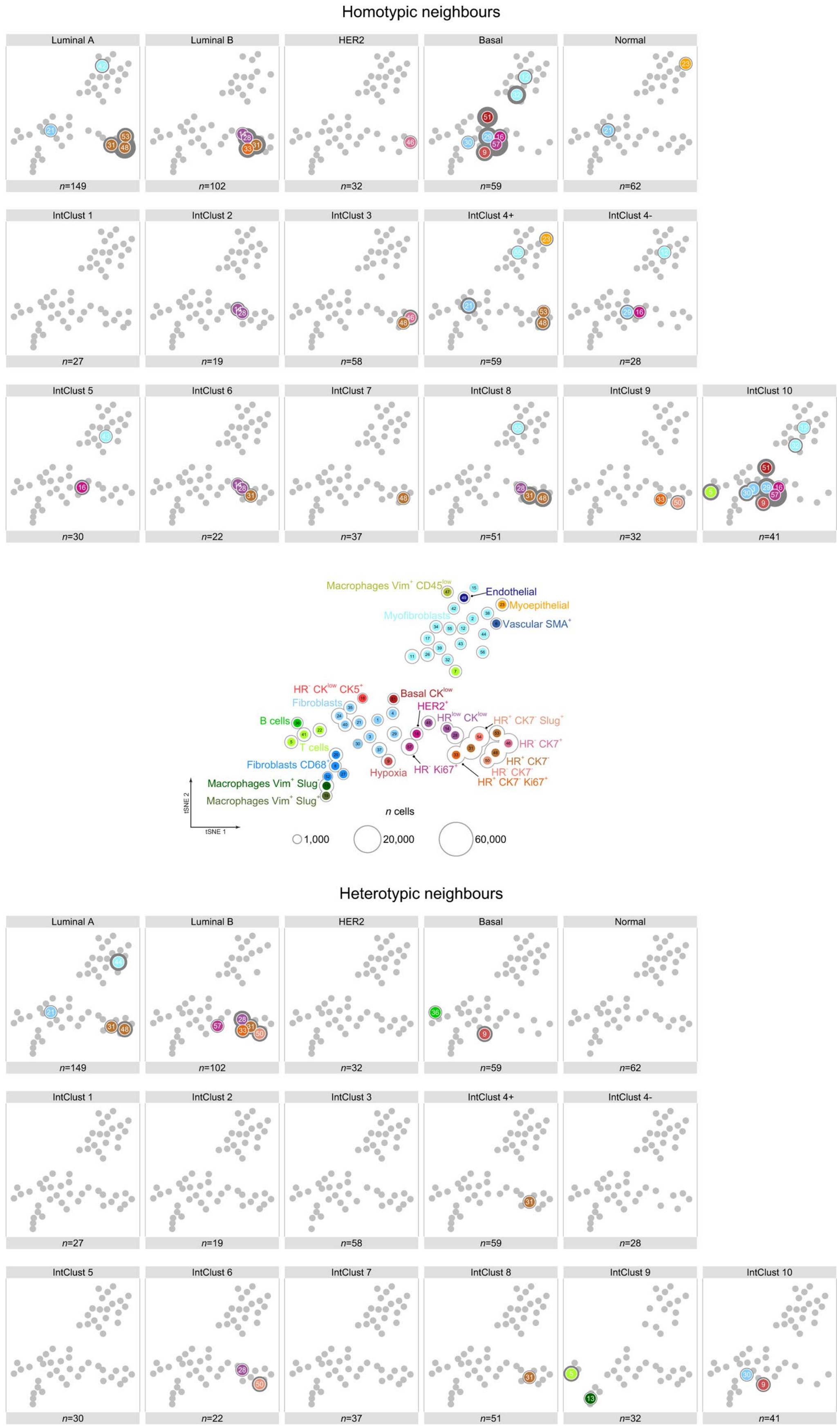


Extended Data Fig. 4 | Robustness of cell phenotype associations with tumour subtypes. a, Example heatmap of Spearman's correlation statistics between 100 FlowSOM groups and Phenograph clusters based on median protein expression values, to illustrate the methodology for mapping FlowSOM groups to cell phenotypes assigned using Phenograph. Solid squares indicate largest positive correlations. Rows and columns are ordered by hierarchical clustering using Ward's method (comparison based on *n* = 479, 844 cells). **b**, Bar chart on the left shows the number of cells assigned to the Phenograph cell phenotypes. Bar chart on the right depicts frequency of FlowSOM groups mapped to each Phenograph-clustered phenotypes, arising from 100 runs of FlowSOM each generating 100 groups (10,000 groups mapped in total). This illustrates that most phenotypes were assigned by both methods and that the frequency of mapped groups was related to the number of cells represented by each phenotype. **c**, Bar charts comparing patterns of association (tested using linear regression) with breast cancer molecular subtypes between mapped cell phenotypes (from 100 runs of FlowSOM each generating 100 groups as in panel b) and the Phenograph-assigned phenotypes. Green and red bars distinguish between enrichment and depletion of a given cell phenotype for each molecular subtype. 'Original' on the x-axis indicates associations based on the Phenograph clustering methodology.



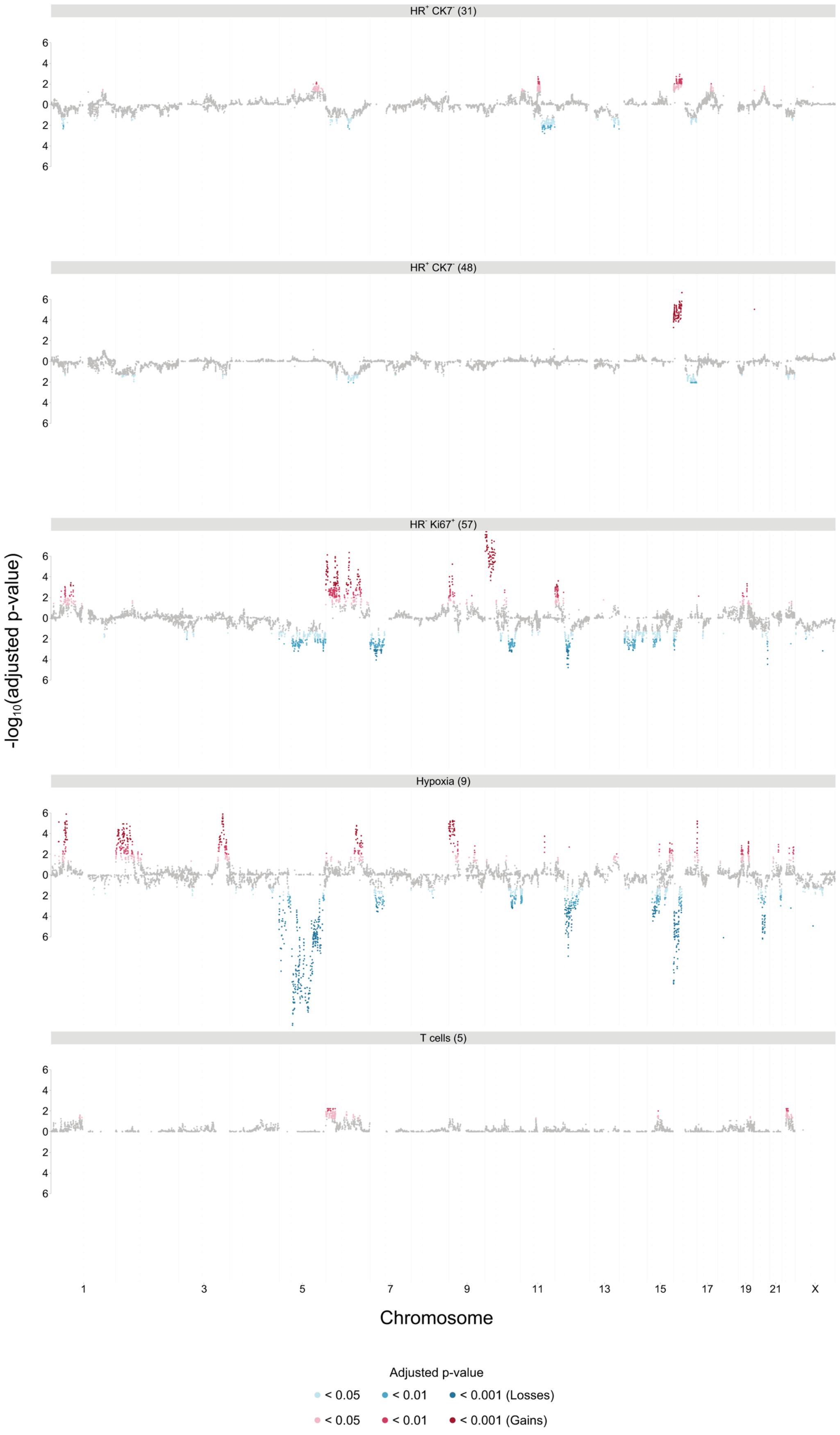
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Patterns of cell phenotype and cell-cell interaction enrichment among breast cancer molecular subtypes. a, Patterns of enrichment among metaclusters defined by combining subsets of Phenograph-assigned cell types by descriptive label (for example, fibroblasts, myofibroblasts, T cells). tSNE map on the right indicates position by cell identity (median values for 22 metaclusters computed based on *n* = 479, 844 cells). Coloured markers indicate significant enrichment. Depicted associations derived from linear regression (*n* = 390 tumours; two-sided tests; adjusted for multiple comparisons), restricted to those with an adjusted p-value < 0.05, were identified by linear models where the dark grey background is proportional to the derived point estimate, providing an indication of the relative strength of the association. **b**, Co-occurrence plots of cell-cell interactions identified by permutation testing and found to be significantly enriched (p-value < 0.05 after adjustment for multiple comparisons) among the molecular subtypes indicated (limited to samples that contain both cells for a given interaction; range of *n* between 86 and 361 tumours for depicted associations). Rows and columns correspond to cell types in the same order as labelled on the y-axis.



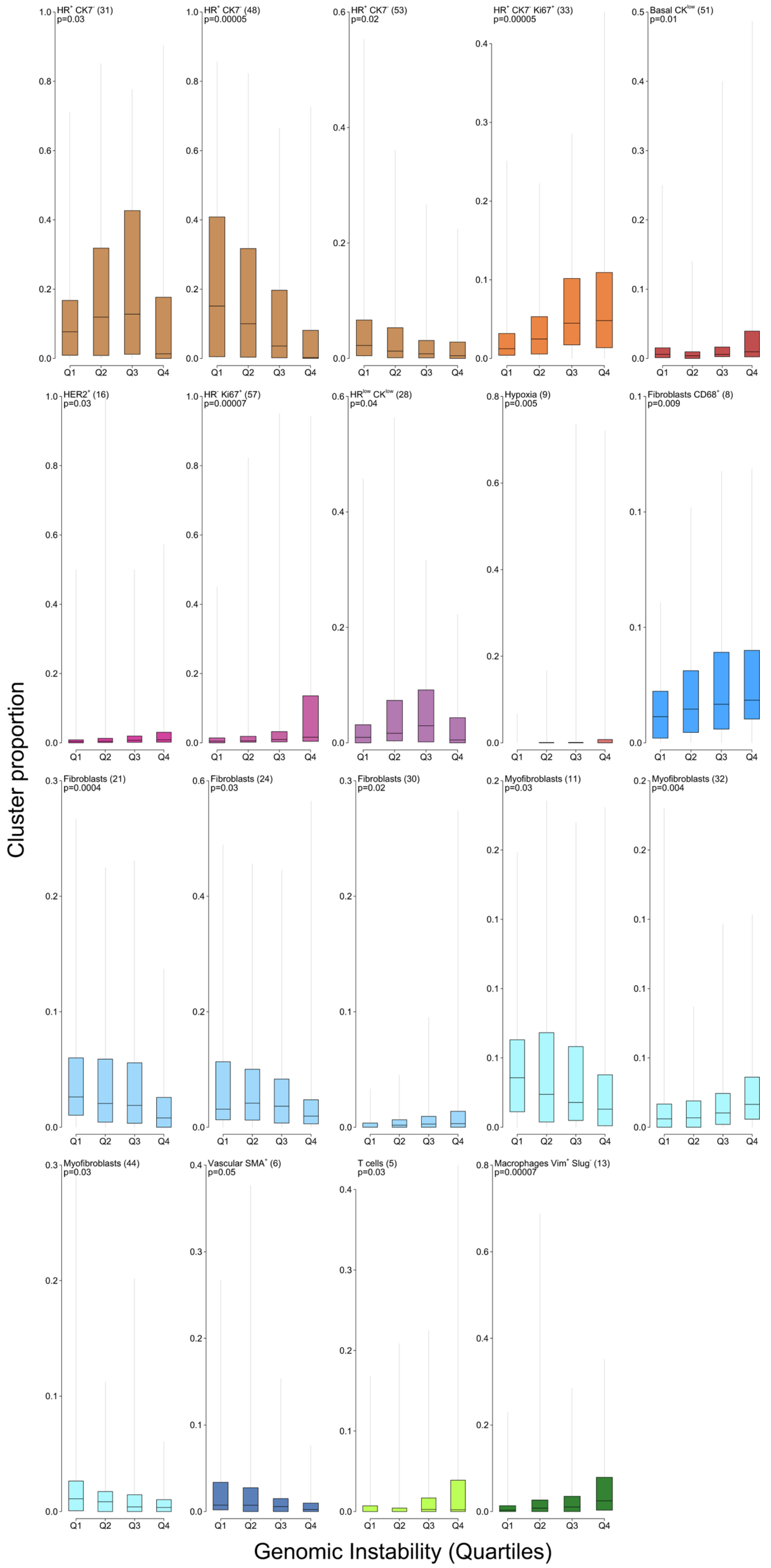
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Distribution of cell neighbourhoods by molecular subtypes of breast cancer. tSNE reference maps representing cell types as light grey makers (tSNE map based on median marker expression values derived from $n=479,844$ cells). Neighbourhood enrichment for each cell type within tumour molecular subtypes was determined by fitting a linear model taking mean neighbourhood values as the dependent variable and tumour molecular subtype as independent variable. Coloured markers indicate those significantly enriched within a given subtype ($p\text{-value} < 0.05$). Dark grey background is proportional to the point estimate from the linear model, providing an indication of the strength (size) of the association.



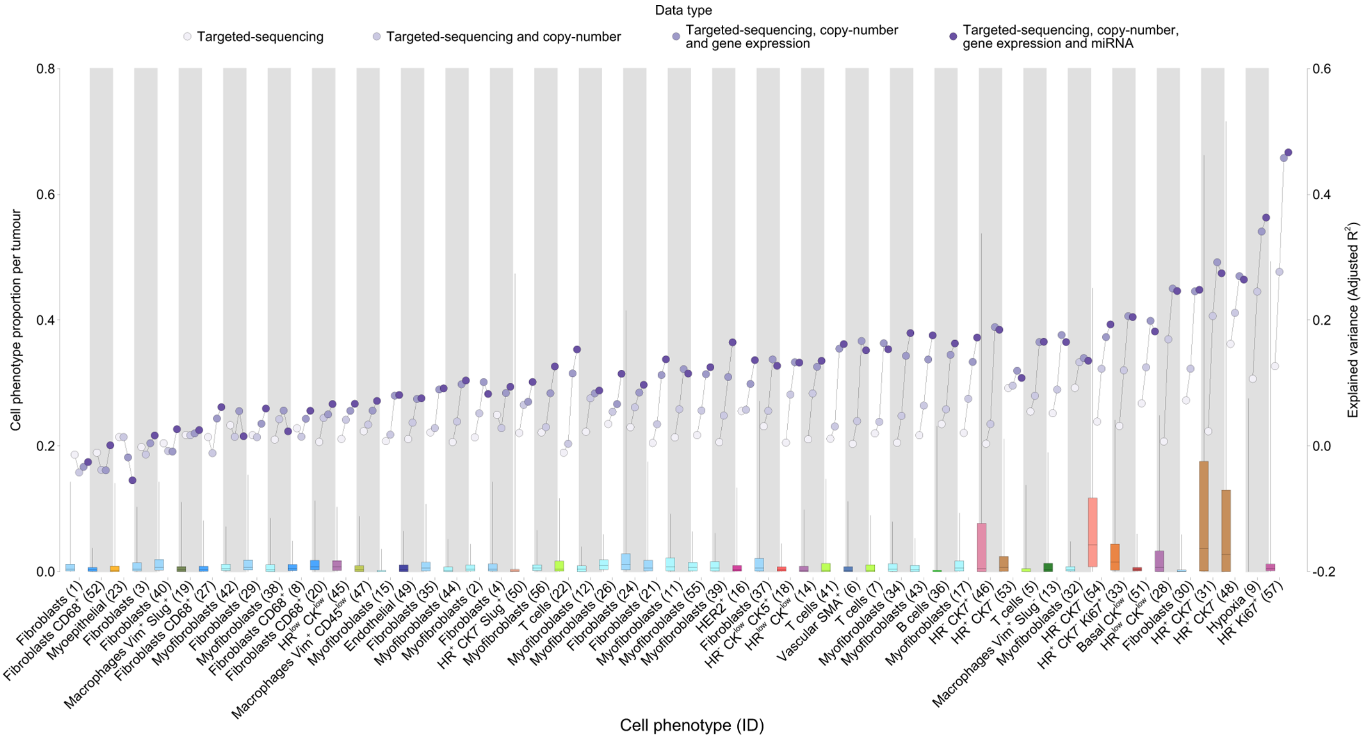
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Associations between cell phenotypes and genome-wide copy number aberrations. Scatter plots depicting adjusted p-values derived from linear models testing for associations between cell phenotype and genomic alterations (*n* = 390 tumours; two-sided; adjusted for multiple comparisons). Coloured points represent significant associations (red, gains; blue, losses). Depicted points are restricted to those associated with positive coefficients. Shown are cell phenotypes most affected by copy-number aberrations.

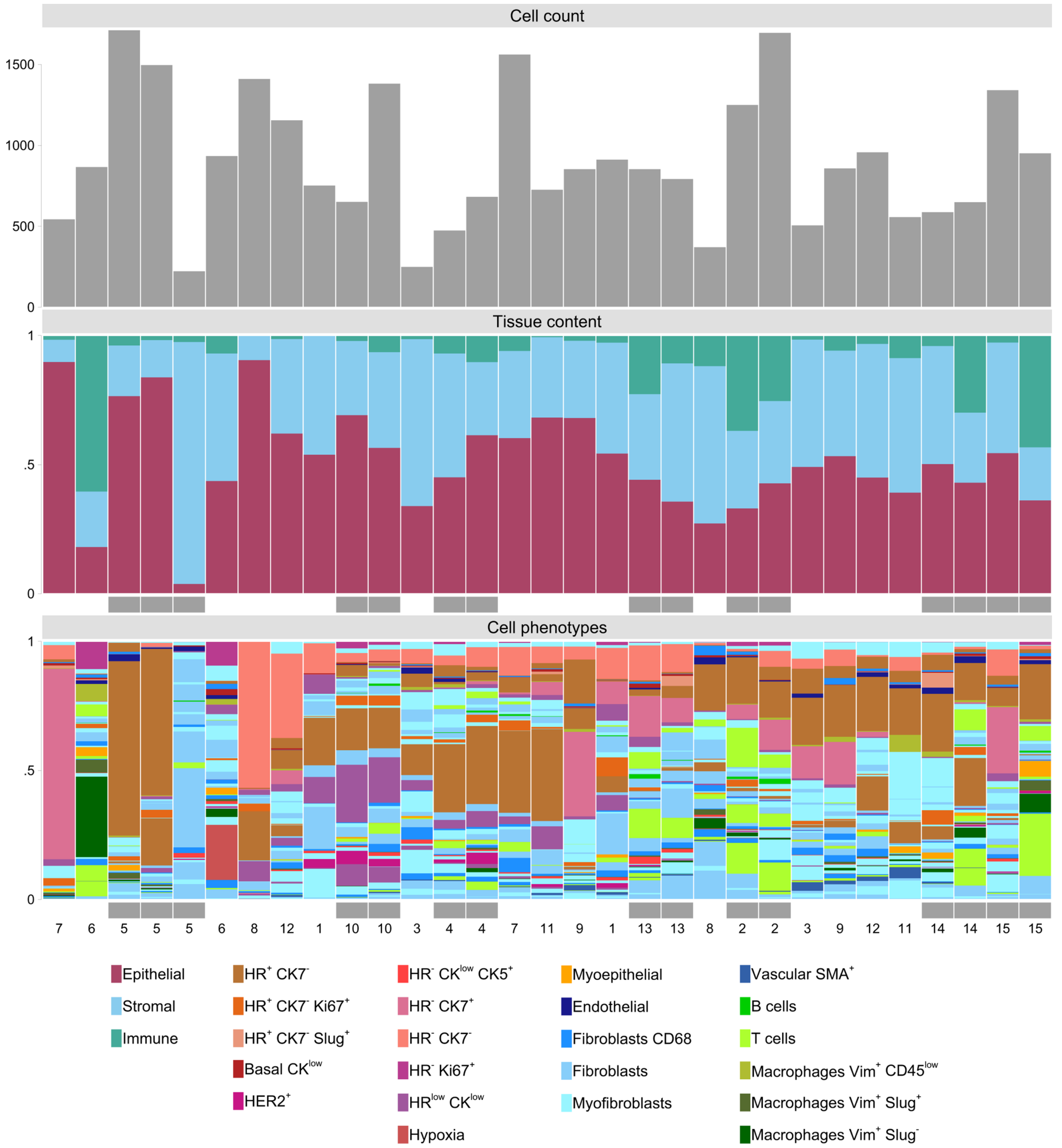


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Relationship between genomic instability and cell type. Box and whisker plots of the distributions of cell types as proportions versus quartiles of genomic instability. Boxes represent the interquartile range. Lines dividing boxes indicate the median, and vertical lines represent range of expression from the 1st to 99th percentile. The p-values were derived from two-sided Kruskal-Wallis tests; depicted are those cell phenotypes with a p-value < 0.05 (adjusted for multiple comparisons; *n* = 404 tumours).



Extended Data Fig. 9 | Cell type abundance explained variance by genomic data. Explained variances (right y-axis) for each of a series of four linear models are depicted as connected circles ($n = 357$ tumours). Distributions of cell type proportions per tumour (left y-axis) depicted as boxes and whiskers. Boxes represent the interquartile range. Lines dividing boxes indicate the median, and vertical lines represent range of expression from the 1st to 99th percentile.



Extended Data Fig. 10 | Influence of tissue sampling on cell phenotype estimates. Stacked bar plot depicting cell phenotype composition per tissue-microarray spot for a subset of fifteen tumours represented by at least two spots. The order of columns was determined using single-linkage hierarchical clustering. Patient IDs are on the x axis. Grey bars highlight where two tissue spots from the same tumour cluster together.

nature research

Corresponding author(s): Bernd Bodenmiller and Carlos Caldas

Last updated by author(s): 19Dec2019

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection

Fluidigm CyTOF Software

Data analysis

In house image preprocessing scripts: https://github.com/BodenmillerGroup/imctools
Ilastik 1.2.0
Cellprofiler version 2.2
Stata SE Version 14.2
R Version 3.4.3
R packages:
clusterprofiler 3.10.1
flowcore 1.48.1
flowsom 1.14.1
reactomepa 1.26.0
glmnet 2.0_16
rphenograph 0.99.1
rtsne 0.15

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

Data

Policy information about availability of data

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

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

Imaging mass cytometry data, including cell masks and processed single cell data, have been deposited to the Image Data Resource (https://idr.openmicroscopy.org/) under the accession code idr0076.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	Sample size was determined by a combination of sample availability and feasibility of assay throughput. Sample size was considered sufficient to represent the breadth of disease (all subtypes) and for survival analyses since standard prognostic factors showed expected effects.
Data exclusions	Samples comprising fewer than 100 segmented cells were excluded from tumour-level analyses owing to insufficient representation of tumour characteristics, as detailed in the manuscript. Criteria were not pre-established.
Replication	Analysis of observational population data; further replication analyses not conducted.
Randomization	Randomisation not relevant to the analysis of observational data.
Blinding	All data acquisition was conducted blinded to associated clinical/molecular data.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study	
□	☒	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
□	☒	Human research participants
□	☒	Clinical data

Methods

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

Antibodies

Antibodies used

Antibodies are detailed in Supplementary Table 1 which includes the following information:
Antigen Antibody Clone Concentration Host Species Clonality Catalogue Number Lot number Vendor
Estrogen Receptor Alpha SP1 8 ug/mL Rabbit Monoclonal M3011* 131011 Spring Bioscience currently Thermo Fisher
Estrogen Receptor Alpha EP1 8 ug/mL Rabbit Monoclonal IR084* AC-0015EU Epitomics, currently Agilent
Histone H3 D1H2 8 ug/mL Rabbit Monoclonal 4499 12 Cell Signaling Technology
Tri-Methyl-Histone H3 (Lys27) C36B11 8 ug/mL Rabbit Monoclonal 9733 11 Cell Signaling Technology
Cytokeratin 5 EP1601Y 6 ug/mL Rabbit Monoclonal ab52635 GR299320-I AbCam
Fibronectin 10/Fibronectin 4 ug/mL Mouse Monoclonal 610078 6251888 BD Biosciences
Cytokeratin 19 Troma-III 6 ug/mL Mouse Monoclonal Troma-III 3/7/13 Dev Studies Hybridoma Bank
Cytokeratin 8/18 C51 4 ug/mL Mouse Monoclonal   4546 2 Cell Signaling Technology
Twist1 Polyclonal 8 ug/mL Rabbit Polyclonal ABD29 2754704 Merck Millipore
CD68 KP1 8 ug/mL Mouse Monoclonal 14-0688-80 E15987-105 eBioscience (ThermoFisher)

Cytokeratin 14 Polyclonal 1 ug/mL Rabbit Polyclonal PA5-16722 SE2391461H eBioscience (ThermoFisher)
SMA 1A4 2 ug/mL Mouse Monoclonal ab7817 033M4768 AbCam
Vimentin D21H3 1 ug/mL Rabbit Monoclonal 5741 3 Cell Signaling Technology
c-Myc 9E10 8 ug/mL Mouse Monoclonal 626802 B201394 Biolegend
c-erbB-2 3B5 8 ug/mL Mouse Monoclonal 554299 6182799 BD Biosciences
CD3ε D7A6E 8 ug/mL Rabbit Monoclonal 85061 2 Cell Signaling Technology
Histone H3 Phospho (Ser28) HTA28 6 ug/mL Rat Monoclonal 641002 B200946 Biolegend
ERK1/2 (pT202/pY204) 20A 8 ug/mL Rabbit Monoclonal 612359 2153932 BD Biosciences
Slug 666633 8 ug/mL Rabbit Monoclonal MAB7408 CFZB021603A R&D Systems
Unconjugated Goat Anti-Rabbit IgG Antibody Polyclonal 4 ug/mL Goat Polyclonal AI-1000 ZA0911 Vector laboratories
Progesterone Receptor A/B EP2 8 ug/mL Rabbit Monoclonal AC-0028EU* N/A Epitomics, currently Abcam
Progesterone Receptor A/B SP2 8 ug/mL Rabbit Monoclonal M3024 131017 Spring Bioscience, currently Abcam
p53 7F5 8 ug/mL Rabbit Monoclonal 2527 Custom purification Cell Signaling Technology
CD44 Polyclonal 6 ug/mL Sheep Polyclonal AF3660 CFOE0216011 R&D Systems
EpCAM (CD326) 9C4 8 ug/mL Mouse Monoclonal 324208 B190626 Biolegend
CD45 2B11 8 ug/mL Mouse Monoclonal 14-9457-80 4298126 eBioscience (ThermoFisher)
GATA3 L50-823 8 ug/mL Mouse Monoclonal 558686 6132744 BD Biosciences
CD20 L26 8 ug/mL Mouse Monoclonal 14-0202-80 4285442 eBioscience (ThermoFisher)
Non-phospho (Active) β-Catenin (Ser33/37/Thr41) (D13A1) D13A1 8 ug/mL Rabbit Monoclonal 8814 2 Cell Signaling Technology
Carbonic Anhydrase IX Polyclonal 8 ug/mL Goat Polyclonal AF2188 VNQ0215032 R&D Systems
E-Cadherin / P-Cadherin 36/E-Cadherin 8 ug/mL Mouse Monoclonal 610181 5274693 BD Biosciences
Ki-67 8D5 3 ug/mL Mouse Monoclonal 9449 2 Cell Signaling Technology
EGF Receptor D38B1 8 ug/mL Rabbit Monoclonal 4267 13 Cell Signaling Technology
Phospho-S6 Ribosomal Protein (Ser235/236) (D57.2.2E) D57.2.2E 6 ug/mL Rabbit Monoclonal 4858 10 Cell Signaling Technology
Sox9 Polyclonal 6 ug/mL Rabbit Polyclonal AB5535 2894178 Merck Millipore
von Willebrand Factor Polyclonal 6 ug/mL Rabbit Polyclonal AB7356 2700933 Merck Millipore
CD31 HC1/6 8 ug/mL Mouse Monoclonal M0823 2557591 Merck Millipore
mTOR, phospho (Ser2448) 49F9 8 ug/mL Rabbit Monoclonal 2976 5536BF Cell Signaling Technology
Cytokeratin 7 RCK105 8 ug/mL Mouse Monoclonal 550507 6083676 BD Biosciences
pan Cytokeratin AE1 0.5 ug/mL Mouse Monoclonal MAB1612 2341224 Merck Millipore
Keratin Epithelial AE3 0.5 ug/mL Mouse Monoclonal MAB1611 2607604 Merck Millipore
Cleaved PARP (Asp214) F21-852 8 ug/mL Mouse Monoclonal 558576 2150663 BD Biosciences
Cleaved Caspase3 C92-605 8 ug/mL Rabbit Monoclonal 559565 559565 BD Biosciences

Validation	Expected antibody staining patterns were confirmed in a variety of tissues including invasive breast carcinoma. Details of vendor validation and use in previous publications are provided as hyperlinks to https://antibodyregistry.org pages in Supplementary Table 1.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Women diagnosed with primary breast cancer between 1985 and 2005.
Recruitment	Selected as a retrospective representative case series; no specific exclusion criteria or patient self-selection were applied.
Ethics oversight	Addenbrooke's Hospital; Cambridge University Hospitals

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

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	The study is not a clinical trial.
Study protocol	The study did not involve a clinical intervention.
Data collection	Tertiary hospital setting; ongoing data collection/updates conducted from approximately 2007 to date.
Outcomes	No specific outcome included in study design.