Imaging mass cytometry and multi-platform genomics define the phenogenomic landscape of breast cancer

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

Author Affiliations

[1]Department of Quantitative Biomedicine, University of Zürich, Zürich, 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]Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, Canada

[5]These authors contributed equally.

*Co-corresponding authors: carlos.caldas@cruk.cam.ac.uk; bernd.bodenmiller@imls.uzh.ch

Abstract

Genomic alterations shape cell phenotypes and the structure of tumour 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 tumours 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 tumour ecosystems, including cellular neighbourhoods, 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 tumour 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-tumour heterogeneity in breast cancer have identified tumour subtypes associated with distinct clinical behaviours[1-3] and driver genomic alterations[4-6]. However, these classifications do not account for the cellular complexity of solid tumours, which are comprised of 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 tumour ecosystems should enable identification of biomarkers and development of new treatments. Here we studied, in unprecedented detail, how genomic alterations shape breast tumour ecosystems by coupling imaging mass cytometry[8] (IMC) to multi-platform genomics. We quantified the abundances of 37 markers in 483 breast tumour samples from the METABRIC cohort[2,5,9], enabling a systematic 'phenogenomic' analysis of breast cancer.

Results

Spatially Resolved Phenotyping of Breast Tumour Ecosystems by IMC

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

We analysed the resulting data using an image processing pipeline adapted for IMC[10-12]. Briefly, we used random forest classification to segment single cells, 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], characterised these breast tumours with unprecedented depth, linking multidimensional tumour phenotypes with somatic genomic alterations.

Data-Driven Derivation of Cell Phenotypes

To investigate cellular diversity and intercellular relationships in breast tumours, we analysed IMC-derived single-cell expression data using a combination of clustering approaches (Fig. 1c). The resulting cell phenotypes fell broadly into the categories of tumour, 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 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, consistent with previous reports[15]. Comparable stromal diversity in breast cancer has recently been reported[16]. For epithelial phenotypes, key distinguishing features included expression of hormone receptors (HRs); cytokeratins 5, 7, and 19; 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 tumour. 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: (i) a group of related cell-cycle pathways active in epithelial cells, (ii) genes necessary for formation of the extracellular matrix and collagen deposition, enriched among myofibroblasts, and (iii) a group of genes related to antigen presentation, interferon gamma signalling, 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 based on IMC data.

miRNAs are critical regulators of cell phenotypes within tumours[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 organisation 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 Characterised by Diverse Tumour Ecosystems

We next compared cell phenotype distributions and spatial features between breast cancer subtypes using linear regression. We focused on two widely used molecular taxonomies of breast cancer: the intrinsic molecular subtypes[1], based on tumour 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 tumour 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 tumours were enriched for HR^+ epithelial cells (phenotypes 31, 48, and 53), whereas more proliferative Luminal B tumours[1] were enriched for both HR^+ epithelial cells (phenotype 31) and HR^+ $Ki67^+$ cells (phenotype 33). Basal-like tumours, which are mostly triple-negative, showed enrichment of HR^- $Ki67^+$ cells (phenotype 57), epithelial cells expressing basal cytokeratins (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 tumours, which map to the Basal-like subtype, showed a near-identical cell enrichment pattern to Basal-like tumours. As expected epithelial cells characterised by high expression of HER2 (phenotype 16) were enriched among the HER2 subtype and IntClust 5 tumours, defined by *ERBB2* amplification.

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

IntClusts 3, 4+, 7, and 8 were all characterised 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 tumours were characterised by the 8p12/*ZNF703* amplicon and IntClust 8 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 tumours. IntClust 3 tumours were characterised by few copy-number alterations, frequent mutations of *PIK3CA*, *CDH1*, and *RUNX1* and the most favourable prognosis of all Integrative Clusters. Enrichment for cell phenotype 46, which showed a highly distinctive expression profile, may indicate that the founding cell of these tumours occupies a different place in the mammary epithelial developmental lineage[18] than founders of other IntClust tumour types.

We observed distinct patterns of stromal cell enrichment in different cancer subtypes (Fig. 4). Fibroblasts that expressed CD68 were enriched among poorer prognosis ER^+ Luminal B tumours. Myofibroblasts were enriched in indolent ER^+ tumours (Luminal A and IntClust 3 and 4+), which are characterised by favourable prognosis but distinct genomic landscapes. Myofibroblasts were also

enriched in IntClust 1 tumours, which are defined by the 17q23 amplicon. The enrichment patterns differed: IntClust 3 tumours, which harbour few CNAs but frequent mutations of *PIK3CA* and *CDH1*, were enriched for myofibroblast phenotype 38 and vascular smooth muscle cell phenotype 6. IntClust 4+ tumours, defined by few genomic aberrations, were enriched for cells of myofibroblast phenotype 55. Fibroblast phenotypes also showed distinct enrichment patterns among indolent ER^+ tumours: Luminal A tumours were enriched for three fibroblast phenotypes (21, 24, and 35), whereas IntClust 3 and 4+ tumours 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 tumours, 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 tumours. This may be related to the high mutational burden, genomic instability, and frequent *TP53* mutations associated with IntClust 10 tumours[5]. Luminal B tumours of the IntClust 9 group were the only ER^+ subtype characterised by both macrophage and T cell enrichment. IntClust 9 tumours have an intermediate-to-poor prognosis, are characterised by 8q amplification, and have the highest proportion of *TP53* mutations among ER^+ tumours[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 tumour subtypes were not adversely affected by signal bleed through or choice of clustering method used to identify cell phenotypes (Extended Data Fig. 3 – 5a).

Genomic tumour subtypes were also characterised 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 tumour subtypes. We distinguished between those of the same cell phenotype (homotypic neighbours) and those of different phenotypes (heterotypic neighbours). Subtypes significantly enriched for interactions included HER2, Basal-like, and IntClust 10 (Extended Data Fig. 5b). In Basal-like and IntClust 10 tumours, we observed abundant homotypic relationships among both epithelial and stromal cells. These tumours are, therefore, distinguished from other subtypes by a starker separation between compartments. We evaluated this further by comparing the average number of homotypic neighbours per cell phenotype and across molecular subtypes (Extended Data Fig. 6). Basal-like and IntClust 10 tumours 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.

Impact of Somatic Genomic Alterations on Breast Tumour Ecosystems

We next investigated associations between cell phenotype and somatic alterations in key driver genes[20]. We compared cell phenotype proportions between tumours with and without a particular alteration

using 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 the $HER2^+$ cell phenotype 16[21]. Similarly, the *TP53* mutation is known to occur more frequently among ER^- tumours than 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 48 and 31) 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^{low} CK^{low}), 31 (epithelial HR^+; lower cytokeratin and hormone receptor expression), and 48 (epithelial HR^+ $CK7^-$) were differentially enriched among Luminal A and B tumours and were associated with distinct genomic events. Cell phenotype 48 was characterised by associations with more mutations than any other phenotype; these included *PIK3CA*, *GATA3*, *MAP3K1*, *CBFB*, *MAP2K4*, *CTCF*, and *MEN1*. In contrast, cell phenotypes 31 and 28 were not associated with mutations, although these phenotypes were associated with CNAs including gains of *CCND1* and *TUBD1* and *ATM* loss. These findings suggest that these ER^+ epithelial cell phenotypes are separated by 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 beta2-microglobulin. This was the only cell phenotype positively associated with both of these alterations. This suggests that tumour cell hypoxia may enable selection of genomic alterations that facilitate immune evasion and supports the previously reported link between tumour 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, 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 tumours, their CNA association profiles diverged substantially. Loss of 5q, a *trans* gene expression module specific to Basal-like tumours 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 tumours, 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 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, tumours 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 proportion is 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 of 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 of myofibroblasts (phenotypes 17, 43, 34, and 39), providing further support that miRNAs are more critical in regulation of gene expression in stromal cells than other cell types in the tumour ecosystem. T cell abundance across all four T cell phenotypes was best explained by gene expression data with little contribution from genomic alterations, consistent with recent work[25].

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

Prognostic Impact of Cell Phenotypes Depends on Their Genomic Context

We examined whether the cell phenotypes and neighbourhoods 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 modelled 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 hypoxia (phenotype 9). Cells within the tumour microenvironment were also prognostic. Macrophages (phenotype 13) were indicative of poorer outcome, whereas vascular smooth muscle cells (phenotype 6) were associated with favourable prognosis. Phenotype 6 cells were enriched among Luminal A and IntClust 3 tumours (Fig. 4).

To assess whether the spatial information in our dataset has 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 tumour 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 tumour microenvironment structure. For example, one group on the heatmap (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 tumour 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 neighbourhoods, where a cell neighbourhood was defined as the cells in contact with a given index cell. We used the mean of homo- or heterotypic cell neighbours per cell phenotype per tumour, normalised to the number of neighbouring cells, for survival analyses (Fig. 6c). Both homo- and heterotypic neighbourhoods showed prognostic associations similar to those of the corresponding cell-proportion predictor. An exception to this trend was the heterotypic neighbourhood 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 neighbourhoods to outcome prediction by fitting a multivariable Cox regression model by penalised maximum-likelihood

estimation. Predictors selected by the model included homo- and heterotypic neighbours in addition to cell proportions (Fig. 6d, e and Supplementary Table 3), suggesting that spatial statistics such as neighbourhoods 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 favourable outcome among IntClust 1 tumours but not among other subtypes. These findings support a model of cancer-associated stroma as a constraint on tumour progression and suggest that this may be related to non-cell autonomous effects of specific genomic alterations.

Cell phenotype 48, characterised by high cytokeratin and HR expression, was also associated with favourable outcome among IntClust 6 tumours but not others. Notably, most IntClust 6 tumours, which are driven by 8p12 amplification, were enriched for tumour cell phenotypes with low cytokeratin and HR expression (phenotypes 31 and 28; Fig. 4). Phenotype 48 cells were characterised by associations with several mutations but not CNAs, contrasting phenotypes 31 and 28 which showed associations with CNAs but not mutations (Fig. 5). Therefore, the subtype-specific prognostic effect of cell phenotype 48 may be related to intra-tumour genetic heterogeneity among IntClust 6 tumours.

The only immune cell phenotype to demonstrate a subtype-specific prognostic effect was phenotype 13 (vimentin$^+$ Slug$^-$ macrophages), which was associated with a favourable outcome among IntClust 7 tumours but with a poorer outcome among other subtypes. IntClust 7 tumours are characterised 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 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 tumour tissue imaging using IMC with multi-platform genomic data to investigate the impacts of somatic alterations on tumour ecosystems at cellular spatial resolution. The tumour samples we studied were from the METABRIC cohort; these samples have been extensively characterised at the genomic level and are linked to long-term patient follow-up data[2,5,9,14]. We quantified the abundances of 37 epitope markers in each sample and used a data-driven approach to phenotypically classify cells and quantify

cellular neighbourhoods 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 favourable-prognosis ER^{+} tumours (Luminal A, IntClusts 3 and 4+) and were characterised by higher cytokeratin and HR expression than cells of phenotypes 28 and 31, which were enriched among poor prognosis luminal tumours (Luminal B and IntClust 6). Most luminal tumours 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 tumours rather than a multimodal distribution[7,29]. Diverse transcriptional programmes regulated by ER lead to the phenotypic diversity in luminal tumours[30,31]. Taken together with our findings, this suggests that the phenotypic compositions of luminal tumours 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 by longitudinal tracking of cell composition, may enable improved clinical decision making.

IntClust 10 tumours, 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 tumours 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 limited 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. Importantly, 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 tumours of all genomic subtypes, but they were most highly enriched in ER^+ tumours with low genomic instability. Survival analysis was suggestive of their associations with favourable outcome (Fig. 6a), in apparent disagreement with the putative pro-tumoural role of CAFs. The myofibroblast phenotype 32 was an exception, as these cells were enriched in IntClust 10 tumours 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 tumours with favourable 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 tumour 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 with myofibroblast phenotype 38 in our analysis. Our findings therefore indicate that a subset of favourable-prognosis luminal breast tumours are characterised by fibroblast activation.

The cardinal features of the multicellular ecosystems of solid tumours have only begun to be explored. Here, integrating multidimensional tissue imaging and multi-platform genomics data for the first time, we identified cellular phenotypic correlates of somatic genomic alterations and demonstrated their variable influence on tumour ecosystems. Our findings suggest that somatic genomic alterations collectively manifest as characteristic tumour ecosystems. Characterisation of these ecosystems will further our understanding of tumour 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.

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. BB's research is funded by a SNSF R`Equip grant, a SNSF Assistant Professorship grant, a 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 n. 336921. The laboratories of B.B. and C.C. were supported by the CRUK IMAX-T Grand Challenge for this work. Members of the CRUK IMAX-T Grand Challenge consortium are:

Ali HR[1,2], Al Sa'd M[3], Alon S[4], Aparicio S[5,6], Battistoni G[1], Balasubramanian S[1,7], Becker R[8], Bodenmiller B[2], Boyden ES[4], Bressan D[1], Bruna A[9], Burger Marcel[2], Caldas C[9], Callari M[1], Cannell IG[1], Casbolt H[1], Chornay N[3], Cui Y[4], Dariush A[3], Dinh K[10], Emenari A[4], Eyal-Lubling Y[9], Fan J[11], Fisher E[1], González-Solares EA[3], González-Fernández C[3], Goodwin D[4], Greenwood W[1], Grimaldi F[8], Hannon GJ[1], Harris O[8], Harris S[8], Jauset C[1], Joyce JA[12], Karagiannis ED[4], Kovačević T[1], Kuett L[2], Kunes R[10], Küpcü Yoldaş A[3], Lai D[5,6], Laks E[5,6], Lee H[11], Lee M[1,7], Lerda G[1], Li Y[5], McPherson A[5,6,13], Millar N[3], Mulvey CM[1], Nugent F[1], O'Flanagan CH[5], Paez-Ribes M[1], Pearsall I[1], Qosaj F[1], Roth AJ[5,6,14], Rueda OM[9], Ruiz T[5], Sawicka K[1], Sepúlveda LA[11], Shah SP[5,6,13], Shea A[9], Sinha A[4], Smith A[5], Tavaré S[1,10,15], Tietscher S[2], Vázquez-García I[13], Vogl SL[8], Walton NA[3], Wassie AT[4], Watson SS[12], Wild SA[1], Williams E[1], Windhager J[2], Xia C[11], Zheng P[11,] Zhuang X[11].

[1] Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge CB2 0RE, UK.

[2] Institute of Molecular Life Sciences, University of Zurich, Zurich 8054, Switzerland.

[3] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge, CB3 0HA, UK.

[4] McGovern Institute, Departments of Biological Engineering and Brain and Cognitive Sciences, Massachusetts Institute of Technology,Cambridge, Massachusetts, USA.

[5] Department of Molecular Oncology, BC Cancer, part of the Provincial Health Services Authority, Vancouver, BC, Canada.

[6] Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada

[7] Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK.

[8] Súil Interactive Ltd, Dame Lane, Dublin, UK.

[9] Department of Oncology and Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, CB2 0RE, UK.

[10] Herbert and Florence Irving Institute for Cancer Dynamics, Columbia University, New York, NY, USA.

[11] Howard Hughes Medical Institute, Department of Physics and of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA.

[12] Department of Oncology and Ludwig Institute for Cancer Research, University of Lausanne, Lausanne, Switzerland.

[13] Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, USA.

[14] Department of Computer Science, University of British Columbia, Vancouver, BC, Canada.

[15] New York Genome Center, New York, NY, USA.

Author Contributions

H.R.A., C.C., and B.B. conceived the study; IMC experiments were performed by H.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 pre-processing 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. 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.

References

1 Perou, C. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747-752, doi:10.1038/35021093 (2000).

2 Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346-352, doi:10.1038/nature10983 (2012).

3 Ali, H. R. *et al.* Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biol* **15**, 431, doi:10.1186/s13059-014-0431-1 (2014).

4 Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nature genetics* **45**, 1127-1133, doi:10.1038/ng.2762 (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, doi:10.1038/ncomms11479 (2016).

6 Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54, doi:10.1038/nature17676 (2016).

7 Wagner, J. *et al.* A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. *Cell* **177**, 1330-1345.e1318, doi:https://doi.org/10.1016/j.cell.2019.03.005 (2019).

8 Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods* **11**, 417-422, doi:10.1038/nmeth.2869 (2014).

9 Dvinge, H. *et al.* The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature* **497**, 378-382, doi:10.1038/nature12108 (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 systems*, doi:10.1016/j.cels.2017.12.001 (2017).

11 Damond, N. *et al.* A Map of Human Type 1 Diabetes Progression by Imaging Mass Cytometry. *Cell metabolism* **29**, 755-768.e755, doi:10.1016/j.cmet.2018.11.014 (2019).

12 Carpenter, A. E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol* **7**, R100, doi:10.1186/gb-2006-7-10-r100 (2006).

13 Haubold, C. *et al.* Segmenting and Tracking Multiple Dividing Targets Using ilastik. *Advances in anatomy, embryology, and cell biology* **219**, 199-229, doi:10.1007/978-3-319-28549-8_8 (2016).

14 Rueda, O. M. *et al.* Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature* **567**, 399-404, doi:10.1038/s41586-019-1007-8 (2019).

15 Gottfried, E. *et al.* Expression of CD68 in non-myeloid cell types. *Scandinavian journal of immunology* **67**, 453-463, doi:10.1111/j.1365-3083.2008.02091.x (2008).

16 Costa, A. *et al.* Fibroblast Heterogeneity and Immunosuppressive Environment in Human Breast Cancer. *Cancer Cell* **33**, 463-479.e410, doi:10.1016/j.ccell.2018.01.011 (2018).

17 Mitra, A. K. *et al.* MicroRNAs reprogram normal fibroblasts into cancer-associated fibroblasts in ovarian cancer. *Cancer Discov* **2**, 1100-1108, doi:10.1158/2159-8290.cd-12-0206 (2012).

18 Stingl, J. & Caldas, C. Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. *Nature Reviews Cancer* **7**, 791, doi:10.1038/nrc2212 (2007).

19 Schapiro, D. *et al.* histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nat Methods*, doi:10.1038/nmeth.4391 (2017).

20 Akavia, U. D. *et al.* An Integrated Approach to Uncover Drivers of Cancer. *Cell* **143**, 1005-1017, doi:10.1016/j.cell.2010.11.013 (2010).

21 Slamon, D. J. *et al.* Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science (New York, N.Y.)* **244**, 707-712, doi:10.1126/science.2470152 (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, doi:10.1016/j.cell.2014.12.033 (2015).

23 Facciabene, A. *et al.* Tumour hypoxia promotes tolerance and angiogenesis via CCL28 and T(reg) cells. *Nature* **475**, 226-230, doi:10.1038/nature10169 (2011).

24 Dawson, S. J., Rueda, O. M., Aparicio, S. & Caldas, C. A new genome-driven integrated classification of breast cancer and its implications. *The EMBO journal* **32**, 617-628, doi:10.1038/emboj.2013.19 (2013).

25 Cristescu, R. *et al.* Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. *Science (New York, N.Y.)* **362**, doi:10.1126/science.aar3593 (2018).

26 Bodenmiller, B. Multiplexed Epitope-Based Tissue Imaging for Discovery and Healthcare Applications. *Cell systems* **2**, 225-238, doi:10.1016/j.cels.2016.03.008 (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, doi:10.1093/annonc/mdu191 (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, doi:10.1371/journal.pmed.1002194 (2016).

29 Reis-Filho, J. S. & Pusztai, L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet* **378**, 1812-1823, doi:10.1016/s0140-6736(11)61539-0 (2011).

30 Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389-393, doi:10.1038/nature10730 (2012).

31 Patten, D. K. *et al.* Enhancer mapping uncovers phenotypic heterogeneity and evolution in patients with luminal breast cancer. *Nature medicine* **24**, 1469-1480, doi:10.1038/s41591-018-0091-x (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, doi:10.1158/0008-5472.can-14-2598 (2014).
33 Bristow, R. G. & Hill, R. P. Hypoxia and metabolism. Hypoxia, DNA repair and genetic instability. *Nat Rev Cancer* **8**, 180-192, doi:10.1038/nrc2344 (2008).
34 Sade-Feldman, M. *et al.* Resistance to checkpoint blockade therapy through inactivation of antigen presentation. *Nat Commun* **8**, 1136, doi:10.1038/s41467-017-01062-w (2017).
35 Kalluri, R. & Zeisberg, M. Fibroblasts in cancer. *Nat Rev Cancer* **6**, 392-401, doi:10.1038/nrc1877 (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, doi:10.1016/j.ccr.2014.04.005 (2014).
37 Rhim, A. D. *et al.* Stromal elements act to restrain, rather than support, pancreatic ductal adenocarcinoma. *Cancer Cell* **25**, 735-747, doi:10.1016/j.ccr.2014.04.021 (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, doi:10.1200/jco.2009.23.5051 (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, doi:10.1038/s41523-018-0056-8 (2018).

Figure Legends

Fig. 1 | Workflow to yield highly multiplexed molecular maps of METABRIC tumours by using imaging mass cytometry (IMC). a, Map of samples ordered by availability of data across platforms. Bar chart depicts the number of segmented cells per tumour. Samples comprising fewer than 100 cells (blue bars) were excluded from tumour-level analyses. **b**, Experimental workflow for multiplexed IMC of 37 proteins in breast tumour tissues associated with genomic annotation and clinical data. Tissue microarrays were labelled 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 neighbourhoods quantified. **c**, Schematic of two-stage clustering approach based on a self-organising map and Phenograph.

Fig. 2 | Data-driven derivation of cellular identities reveals composition of tumour ecosystems. a, Two-dimensional tSNE representation of multiplexed proteomic data highlighted by cell phenotype. Each dot represents one cell; 5% of cells per tumour were randomly selected for illustration (n = 24,003 cells*)*. b, tSNE maps coloured by expression of five canonical proteins. c, Bar plot showing the relative proportions of epithelial, stromal, and immune cells of all cells analysed. d, Annotated tSNE map of cell phenotypes drawn using median protein expression levels. e, Heatmap of pairwise Spearman rank correlations of cell proportions, where the total cell count per tumour was taken as the denominator. Cell phenotypes were ordered by hierarchical clustering by Ward's method. Highlighted squares indicate significant cell-cell interactions (determined by permutation tests) observed in at least 10% of tumours. f, Heatmap of median values of normalised protein expression per cell cluster. Markers were arranged by hierarchical clustering by Ward's method. Bar chart on the right depicts total cell count per cluster, distinguishing those cells derived from ER^+ versus ER^- tumours.

Fig. 3 | Transcriptomic correlations with IMC-defined cell types. a, Scatter plot comparing the number of positive and negative correlations between cell phenotype proportions and gene expression levels determined using linear models (n = 390 tumours). Each point represents one cell phenotype (n = 55 cell phenotypes; spearman correlation = 0.95; p-value < 0.05). b, Comparative reactome pathway enrichment analysis by cell phenotype based on the most strongly positively correlated genes (n = 390 tumours; 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. c, Scatter plot comparing the number of positive and negative correlations between cell cluster proportions and miRNA expression levels determined using linear models (n = 371 tumours). Each point represents one cell phenotype (n = 55 cell phenotypes; spearman correlation = 0.58; p-value < 0.05). ECM, extracellular matrix; IFN, interferon; TCR, T cell receptor.

Fig. 4 | Phenotype enrichment in genomic breast cancer subtypes. Enriched phenotypes in each indicated genomic subtype are illustrated as two-dimensional tSNE 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 with a p-value < 0.05 (two-sided, adjusted for multiple comparisons per subtype by Benjamini-Hochberg correction). The dark grey background is proportional to the model coefficient, providing an indication of the strength of the association.

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

Fig. 6 | Prognostic impact of cell phenotypes and their neighbourhoods. a, Hazard ratios of disease-specific survival for each cell type, modelled as log-ratios. Circles represent point estimates and whiskers the 95% confidence interval derived from a Cox proportional-hazards model, adjusted for ER status (n = 448 patients, two-sided tests). b, Heatmap of Spearman rank correlations among cell proportions of each cell type, relative to all cells analysed per tumour (n = 467 tumours). Cell phenotypes are ordered by hierarchical clustering. Highlighted orange and blue squares of cell-cell

relationships are restricted to significant cell-cell interactions (orange) or separation (blue) determined by permutation tests and observed in at least 10% of images. c, Scatter plots comparing hazard ratios of cell neighbourhoods and cell proportions (Cox-regression, n = 448 patients). Highlighted are those with a cell neighbourhood p-value < 0.05 but a cell proportion p-value > 0.05. d, Phenotypes and neighbourhoods selected by a multivariable model as predictors of disease-specific survival (regularised Cox-regression, n = 448 patients). Coloured markers represent features selected by the model with red indicating an association with poorer outcome (greater hazard) and blue an association with better outcome (lesser hazard); precise hazard ratios are provided in Supplementary Table 3. e, Survival plot by quartiles of values (hazard ratio) predicted using 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 with a p-value < 0.05 for interaction between cell phenotype and IntClust subtype (derived from a Cox-regression model adjusted for ER-status, n = 390 patients, p-value adjusted for multiple comparisons).

Methods

Study Population and Genomic Assays

We analysed breast tumour 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-anonymised clinical data. Extensive details of specimen handling, nucleic acid extraction, microarray hybridisation, 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 using 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 was conducted using Affymetrix SNP 6.0 arrays, and transcriptional profiling was conducted using the Illumina HT-12 v3 platform. Segmentation and copy-number calls were made using circular binary segmentation, and gene expression data were normalised using the beadarray[40] R package. miRNA profiling was conducted using a custom Agilent microarray in which putative and known miRNA sequences were represented. For targeted sequencing, libraries were prepared using 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 using 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 haematoxylin and eosin stained slides by a breast pathologist (E.P.). Cores of 0.6 mm corresponding to marked areas were

then removed and processed as previously described[41]. Of the 483 tumours included in this analysis, 463 were represented by one core, 19 by two cores, and one by three cores. Where tumours 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 tumours 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 between paired cores using hierarchical clustering. For seven of the fifteen 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 tumour than between those from different tumours. Where cores from a tumour did not cluster together, this was often because the tissue content differed between them. For example, one contained mostly stromal cells, whereas the other contained mostly tumour cells. Therefore, although the study was not free of sampling error, these observations suggest that it did not represent a major impediment.

Antibody Conjugation

Descriptions of antibodies, isotope tags, and concentrations used for staining are provided in Extended Data Table 1. Antibody-metal conjugation was conducted using the Maxpar labelling kit (Fluidigm). Following conjugation, the concentration was assessed using a Nanodrop (Thermo Scientific) and was adjusted to between 100 and 500 µg/ml. Antibodies were stored in Candor Antibody Stabiliser (Candor Bioscience) at 4 ºC. 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 Labelling

Slides were stained as previously described[19]. Briefly, slides were deparaffinised in xylene and rehydrated in a graded alcohol series. Antigen retrieval was conducted using Tris-EDTA (pH 9) buffer at 95 ºC in a NxGen decloaking chamber (Biocare Medical). Following cooling, slides were blocked with 3% BSA in TBS for 1 hour. Slides were incubated with metal-tagged antibodies overnight at 4 ºC with the exception of anti-oestrogen receptor alpha antibodies, which were detected using a metal-tagged anti-rabbit secondary antibody to increase signal (Extended Data Table 1). 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

Abundance of bound antibody was quantified using 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 Neighbourhoods

Count data were converted to tiff image stacks and analysed using a bespoke image processing pipeline (https://github.com/BodenmillerGroup/imctools). Briefly, random 125 x 125 µm crops of images were generated and up-scaled by a factor of 2 for pixel classification using the pixel-classification tool ilastik. Pixels were manually labelled as nuclear, cytoplasmic, and background to train a random forest classifier in ilastik. The trained classifier was used to attribute probabilities to 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 per image; we then extended the training set using pixels from these images and repeated the process until improvement in model performance plateaued (four iterations). Probability maps were analysed using CellProfiler[12]. Nuclei were detected as primary objects with secondary objects and cytoplasm and cell membrane were identified by expanding primary objects to the border between cell cytoplasm/membrane and background using the propagation method. Single-cell regions identified in this way formed a cell mask used for signal quantification and derivation of neighbourhood 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 using ilastik, generated a corresponding mask and removed affected cells from analyses. We found that the majority of cells from the two rare 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. Processed data will be made available upon publication of the manuscript.

Cell Clustering

Single-cell expression data were arcsinh transformed using 0.8 as a cofactor prior to analysis. Based on protein distribution values across all cells, data were clipped at the 99th centile and cells 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: CK8/18, CK19, CK5, CD68, CD3, CD20, ER, PR, CD45, GATA3, CK7, Ki67, SMA, HER2, panCK, EGFR, TP53, beta-catenin, vWF/CD31, CAIX, Slug, and vimentin. We analysed data in two stages. First, we clustered cells into 225 groups using a self-organising map[44] (15 x 15) implemented in the FlowSOM package[44] and then, using the mean expression values within each of these clusters per image, conducted a second round of clustering using the community detection algorithm Phenograph[45] resulting in 57 clusters (of which two 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 based on markers of epithelial (panCK, CK7, CK8/18,

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 based on cell location and morphology.

Cell-Cell Interactions and Cell Neighbourhoods

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, immediate neighbours of each cell as defined in the 'Object Relationships' table created using the CellProfiler pipeline were used to generate a null distribution of cell interaction frequencies by permuting cell labels 1,000 times per 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 one 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 using the Benjamini-Hochberg method[47,48]. Cell neighbourhood statistics were computed for each tumour as the average number of adjacent homo- or heterotypic neighbours per cell, adjusted for the number of neighbours. Homotypic neighbourhood statistics were computed as the average number of cell neighbours that were of the same cell phenotype, and heterotypic neighbourhoods as the average number that were of a different phenotype.

Statistics and Reproducibility

Cell phenotypes were treated as proportions. Spearman rank correlations were computed based on the proportion of a cell phenotype compared to all of the cells in a tumour. For comparison to genomic and clinical data, cell phenotype proportions were computed separately by whether a cell was epithelial or not epithelial. Adjustments for multiple testing were conducted using the Benjamini-Hochberg method[47,48]. No statistical method was used to predetermine sample size. Samples comprising fewer than 100 cells were removed from tumour-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 tumour subtypes were determined using the PAM50 method as previously described[2,49]. Integrative cluster subtype was based on the original designation[2]. Enrichment of cell phenotypes by molecular subtype was tested separately for each subtype using 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 using 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 tumour subtypes was related to signal bleed from tumour cells into adjacent stromal cells in closely packed areas, where cell segmentation can be problematic. We therefore examined the composition of all neighbouring cells for each cell phenotype (Extended Data Fig. 7a, b). This showed that most neighbouring 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 neighbours than others (Extended Data Fig. 7a). When we compared the composition of neighbouring cells separately for genomic subtypes of breast cancer, we did not find that those identified as enriched were neighboured by a greater proportion of epithelial cells compared to those that were not significantly enriched (Extended Data Fig. 7b). Stromal cell enrichment patterns among tumour subtypes were not, therefore, due to inappropriate attribution of tumour 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 using a machine-learning based pixel classification. In the second strategy, a mask was drawn around each nucleus up to a maximum distance of 3 pixels, not including background, resulting in a shrunken mask per cell. We then compared the expression profiles of stromal cell phenotypes based on either whole-cell or 3-pixel segmentation limited to cells that mapped unambiguously. A clustered heatmap (Extended Data Fig. 7c) revealed that stromal phenotype expression profiles based on whole-cell segmentation clustered together with 3-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 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 tumour subtypes were robust to the cell clustering method used to identify cell phenotypes, we used FlowSOM rather than the clustering strategy using Phenograph to cluster cells into 100 groups. We then mapped these groups to each of the 57 cell phenotypes identified by our original method based on the similarity of their expression profiles (Extended Data Fig. 8a). Finally, we tested for associations between these mapped cell phenotypes and tumour subtypes to compare patterns of association between the original cell phenotypes and their mapped counterparts. To account for random initialisation 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 that most phenotypes were successfully recovered (Extended Data Fig. 8b). There was

excellent concordance for associations with genomic tumour subtypes between the original cell phenotypes and newly mapped groups (Extended Data Fig. 8b). In sum, these findings showed that patterns of association with tumour subtypes were robust to choice of clustering strategy.

We also investigated whether combining all cells belonging to cell phenotypes with the same descriptive label (e.g., fibroblasts) would lead to a meaningful loss of information. We combined cell phenotypes with the same descriptive label and tested for enrichment patterns among tumour 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 tumours, 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 tumour ecosystems of different tumour subtypes.

Correlation of Cell Phenotype with Gene Expression or miRNA Expression

Gene expression and miRNA data, processed and normalised as previously described[2,9], were used for these analyses. Where more than one probe mapped to a gene, the probe with greatest variance across the dataset was selected. Cell phenotype correlations with gene expression and miRNAs were estimated using 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 using 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 using 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 generalised additive model fit to the entire METABRIC cohort as previously described[9]. Pathways enriched among the resulting targets were identified as for gene expression analyses.

Correlation of Cell Phenotype with Genomic Variation

Data processing and normalisation were conducted as previously described[2,5]. Genomic instability was computed as the proportion of the non-diploid genome based on ASCAT integer copy-number calls[52]. Kruskal-Wallis tests were used to test for the 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 tumours were coded as either positive or negative for a given copy number alteration. A similar strategy was used to analyse associations with mutations. Tumours were deemed either positive or negative for a given mutation encompassing all non-synonymous mutations; genes with fewer than five mutations observed were

excluded from association analyses based on data contained in supplementary table 4 of reference[5]. A given CNA or mutation was tested for association with a cell phenotype 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 tumour and the total number of copy-number events per tumour. 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 per 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 using 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-squared statistic as an indicator of explained variance.

Survival Analyses

Analyses were based on updated clinical data available in reference[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 in order 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 modelled 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 using Benjamini-Hochberg correction. Evaluation of all log-ratios and neighbourhoods (163 predictors) in a multivariate model was conducted using a penalised maximum-likelihood estimated Cox regression model implemented in the R package glmnet[56]. Lambda was selected using cross-validation. All analyses were conducted using Stata SE version 14.2 and R[57].

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

References

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, doi:10.1093/bioinformatics/btm311 (2007).

41 Kononen, J. *et al.* Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature medicine* **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, doi:10.1186/s13059-016-1006-0 (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.e1319, doi:10.1016/j.cell.2018.08.039 (2018).

44 Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry. Part A : the journal of the International Society for Analytical Cytology* **87**, 636-645, doi:10.1002/cyto.a.22625 (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, doi:10.1016/j.cell.2015.05.047 (2015).

46 Phipson, B. & Smyth, G. K. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology* **9**, Article39, doi:10.2202/1544-6115.1585 (2010).

47 Newson, R. B. Frequentist q-values for multiple-test procedures. *Stata Journal* **10**, 568-584 (2010).

48 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **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, doi:10.1200/jco.2008.18.1370 (2009).

50 Yu, G. & He, Q. Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular bioSystems* **12**, 477-479, doi:10.1039/c5mb00663e (2016).

51 Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology* **16**, 284-287, doi:10.1089/omi.2011.0118 (2012).

52 Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-16915, doi:10.1073/pnas.1009843107 (2010).

53 Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer*, doi:10.1038/s41568-018-0060-1 (2018).

54 Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in microbiology* **8**, 2224, doi:10.3389/fmicb.2017.02224 (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, doi:10.1371/journal.pmed.1000279 (2010).

56 Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software; Vol 1, Issue 1 (2010)*, doi:10.18637/jss.v033.i01 (2010).

57 R: A language and environment for statistical computing. (https://www.R-project.org/) (R Foundation for Statistical Computing, Vienna, Austria, 2016).

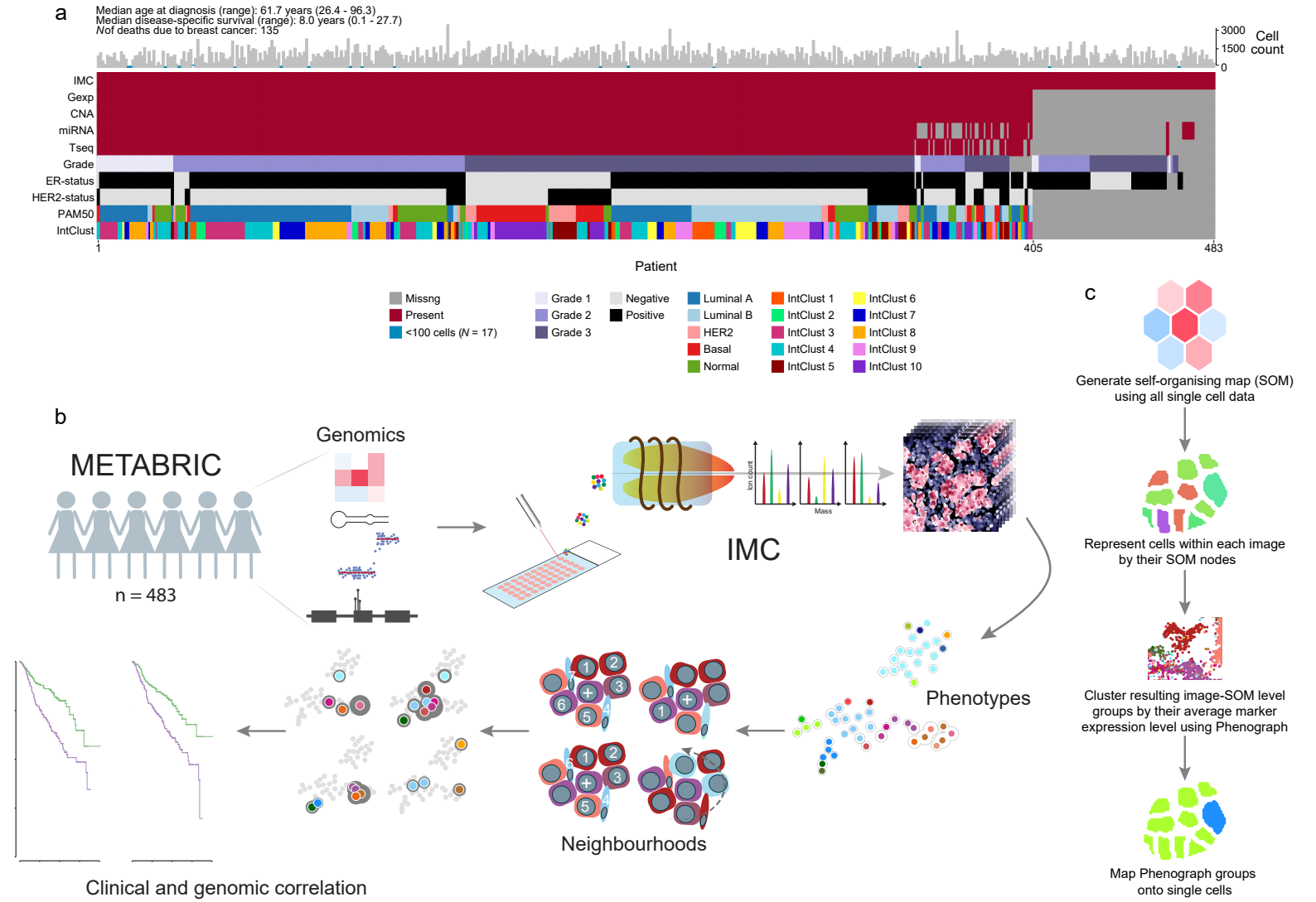
Data Availability

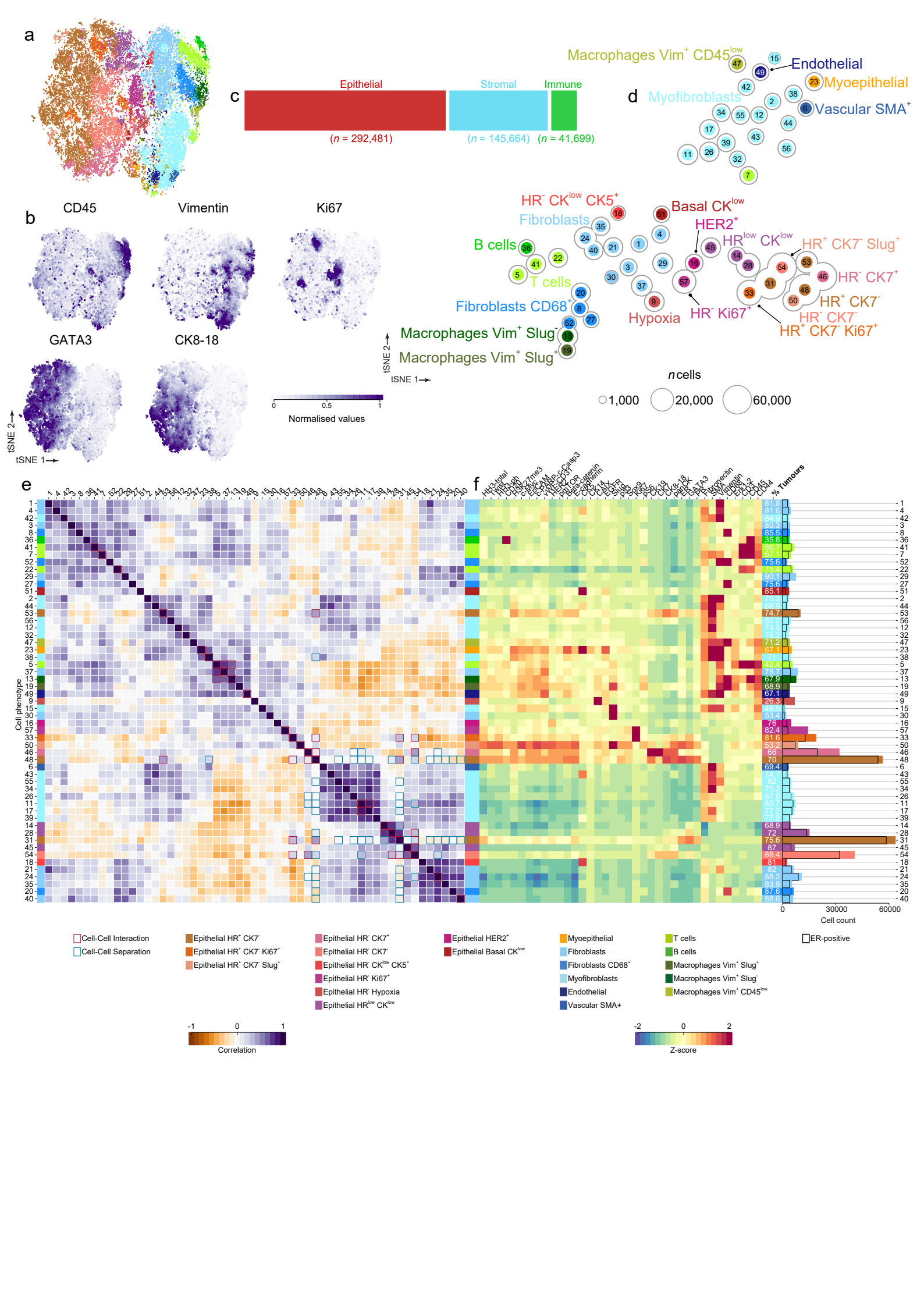
Imaging mass cytometry data, including cell masks and processed single cell data, have been deposited to the Image Data Resource (https://idr.openmicroscopy.org/) under the accession code idr0076. Previously published METABRIC copy-number, gene expression, miRNA and targeted sequencing

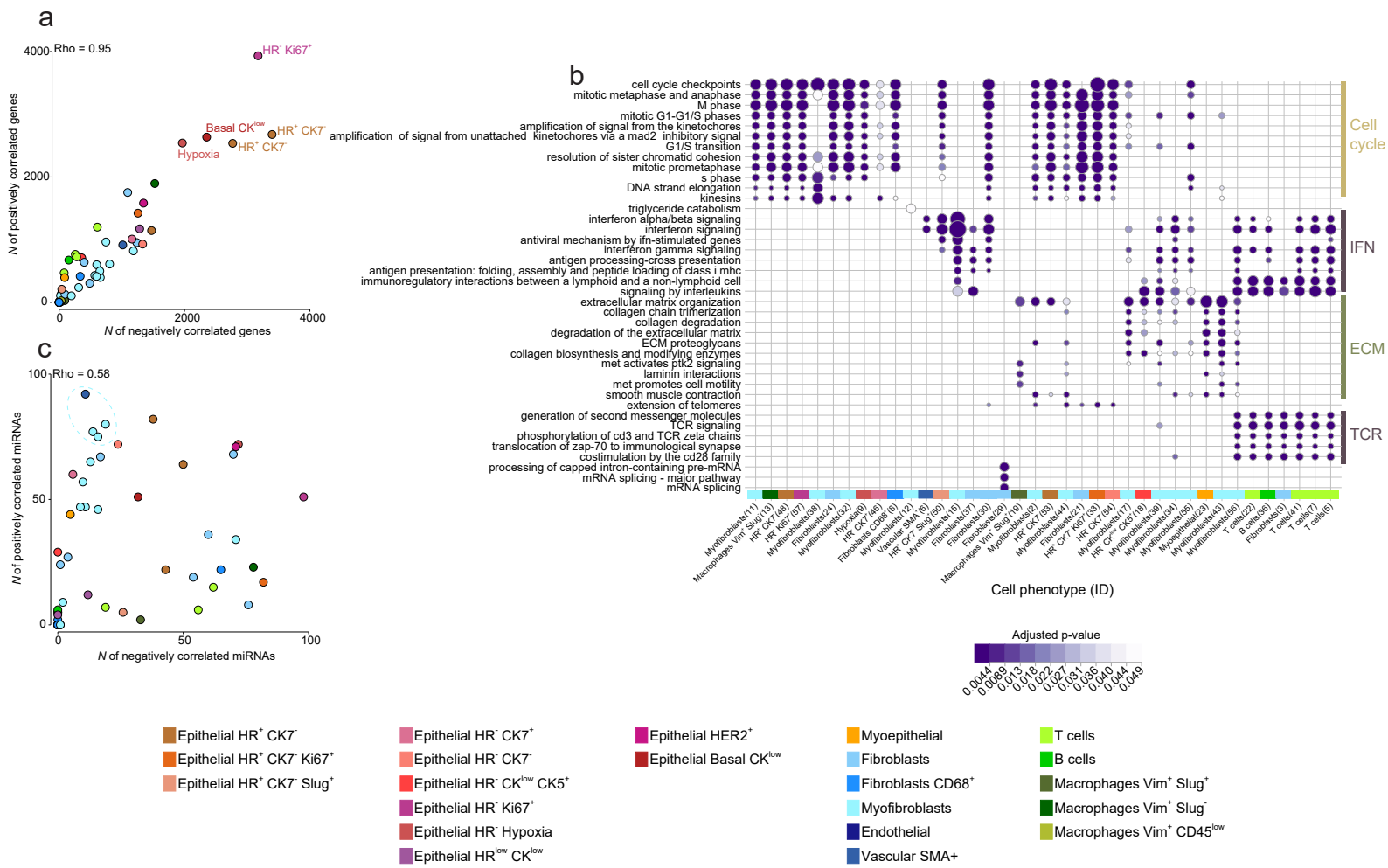
data that were re-analysed here are available under accession codes EGAS00000000083, GAS00000000122 and EGAS00001001753 at the European Genome-Phenome archive (http://www.ebi.ac.uk/ega/). Updated METABRIC clinical data analysed here are available as part of the supplementary information in reference[14]. All other data supporting the findings of this study are available from the corresponding author on reasonable request.

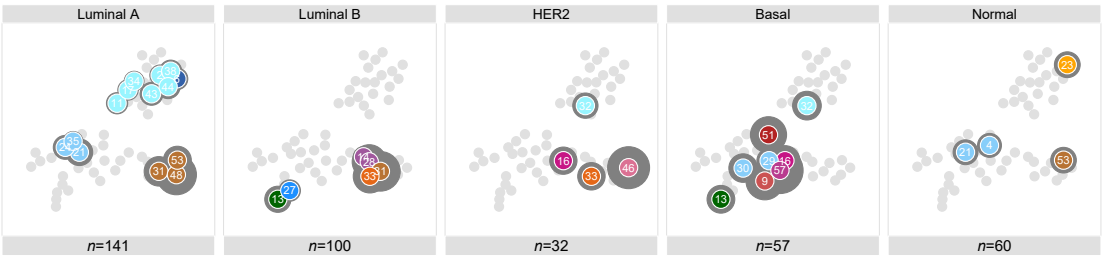
Code Availability

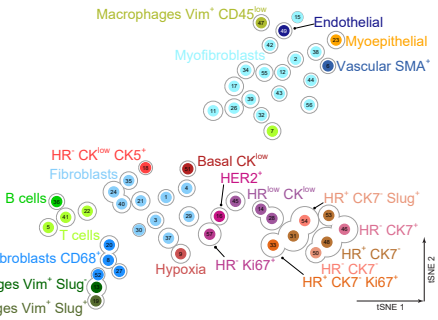
In house image preprocessing scripts are available at https://github.com/BodenmillerGroup/imctools. Other analysis code is available from the authors upon request.

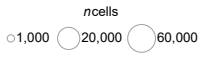

a
Median age at diagnosis (range): 61.7 years (26.4 - 96.3)
Median disease-specific survival (range): 8.0 years (0.1 - 27.7)
Nof deaths due to breast cancer: 135
3000
1500
0
Cell count
IMC
Gexp
CNA
miRNA
Tseq
Grade
ER-status
HER2-status
PAM50
IntClust
1
405
483
Patient
Missng
Present
<100 cells (N = 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
IMC
Phenotypes
Neighbourhoods
Clinical and genomic correlation
c
Generate self-organising map (SOM) using 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 using Phenograph
Map Phenograph groups onto single cells

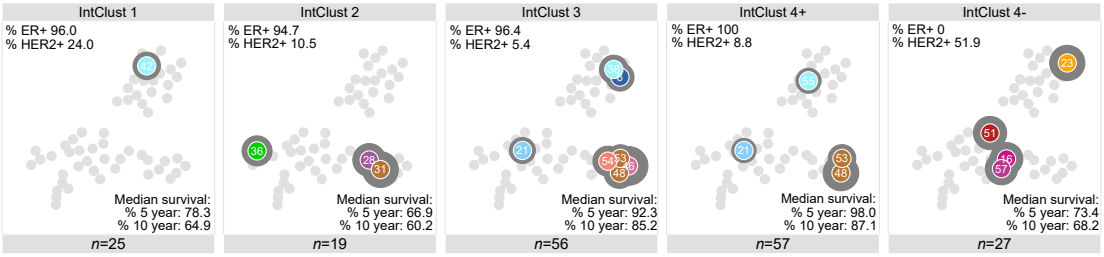

a
b
CD45
Vimentin
Ki67
GATA3
CK8-18
tSNE 2
tSNE 1
0
0.5
1
Normalised values
c
Epithelial
Stromal
Immune
(n = 292,481)
(n = 145,664)
(n = 41,699)
d
Macrophages Vim+ CD45low
Endothelial
Myoepithelial
Myofibroblasts
Vascular SMA+
HR- CKlow CK5+
Basal CKlow
HER2+
Fibroblasts
HRlow CKlow
HR+ CK7- Slug+
B cells
HR- CK7+
T cells
HR+ CK7-
Fibroblasts CD68+
Hypoxia
HR- Ki67+
HR- CK7-
HR+ CK7- Ki67+
Macrophages Vim+ Slug-
Macrophages Vim+ Slug+
n cells
1,000
20,000
60,000
e
f
Cell phenotype
% Tumours
Cell count
0
30000
60000
Cell-Cell Interaction
Cell-Cell Separation
Epithelial HR+ CK7-
Epithelial HR+ CK7- Ki67+
Epithelial HR+ CK7- Slug+
Epithelial HR- CK7+
Epithelial HR- CK7-
Epithelial HR- CKlow CK5+
Epithelial HR- Ki67+
Epithelial HR- Hypoxia
Epithelial HRlow CKlow
Epithelial HER2+
Epithelial Basal CKlow
Myoepithelial
Fibroblasts
Fibroblasts CD68+
Myofibroblasts
Endothelial
Vascular SMA+
T cells
B cells
Macrophages Vim+ Slug+
Macrophages Vim+ Slug-
Macrophages Vim+ CD45low
ER-positive
-1
0
1
Correlation
-2
0
2
Z-score

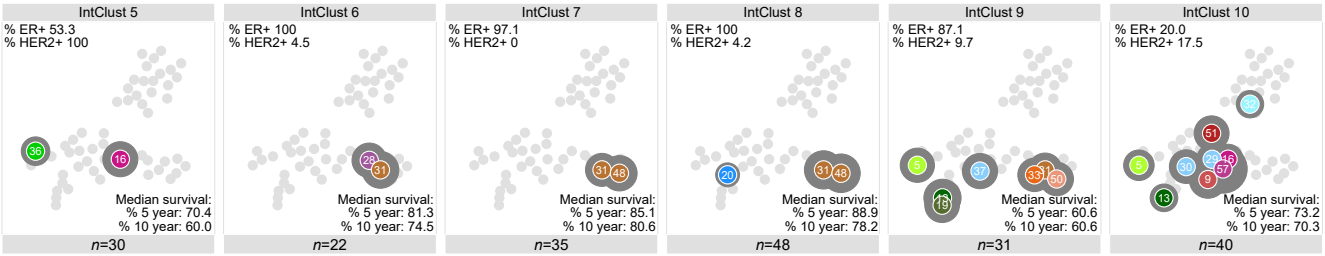

a
Rho = 0.95
HR- Ki67+
Basal CKlow
HR+ CK7-
HR+ CK7-
Hypoxia
N of positively correlated genes
N of negatively correlated genes
b
cell cycle checkpoints
mitotic metaphase and anaphase
M phase
mitotic G1-G1/S phases
amplification of signal from the kinetochores
amplification of signal from unattached kinetochores via a mad2 inhibitory signal
G1/S transition
resolution of sister chromatid cohesion
mitotic prometaphase
s phase
DNA strand elongation
kinesins
triglyceride catabolism
interferon alpha/beta signaling
interferon signaling
antiviral mechanism by ifn-stimulated genes
interferon gamma signaling
antigen processing-cross presentation
antigen presentation: folding, assembly and peptide loading of class i mhc
immunoregulatory interactions between a lymphoid and a non-lymphoid cell
signaling by interleukins
extracellular matrix organization
collagen chain trimerization
collagen degradation
degradation of the extracellular matrix
ECM proteoglycans
collagen biosynthesis and modifying enzymes
met activates ptk2 signaling
laminin interactions
met promotes cell motility
smooth muscle contraction
extension of telomeres
generation of second messenger molecules
TCR signaling
phosphorylation of cd3 and TCR zeta chains
translocation of zap-70 to immunological synapse
costimulation by the cd28 family
processing of capped intron-containing pre-mRNA
mRNA splicing - major pathway
mRNA splicing
Cell cycle
IFN
ECM
TCR
Cell phenotype (ID)
Adjusted p-value
0.0044
0.0089
0.013
0.018
0.022
0.027
0.031
0.036
0.040
0.044
0.049
c
Rho = 0.58
N of positively correlated miRNAs
N of negatively correlated miRNAs
Epithelial HR+ CK7-
Epithelial HR+ CK7- Ki67+
Epithelial HR+ CK7- Slug+
Epithelial HR- CK7+
Epithelial HR- CK7-
Epithelial HR- CKlow CK5+
Epithelial HR- Ki67+
Epithelial HR- Hypoxia
Epithelial HRlow CKlow
Epithelial HER2+
Epithelial Basal CKlow
Myoepithelial
Fibroblasts
Fibroblasts CD68+
Myofibroblasts
Endothelial
Vascular SMA+
T cells
B cells
Macrophages Vim+ Slug+
Macrophages Vim+ Slug-
Macrophages Vim+ CD45low

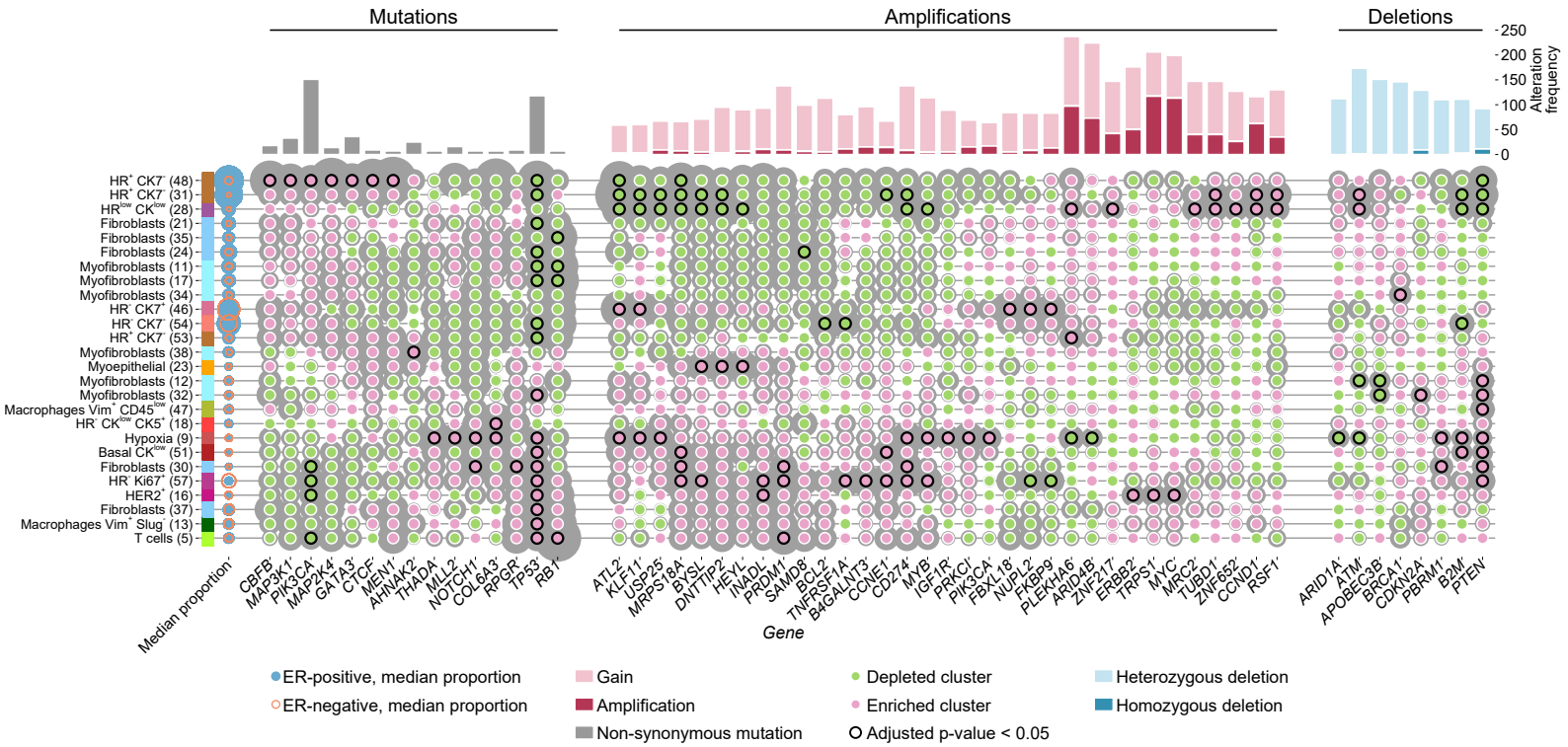

Luminal A
n=141
Luminal B
n=100
HER2
n=32
Basal
n=57
Normal
n=60

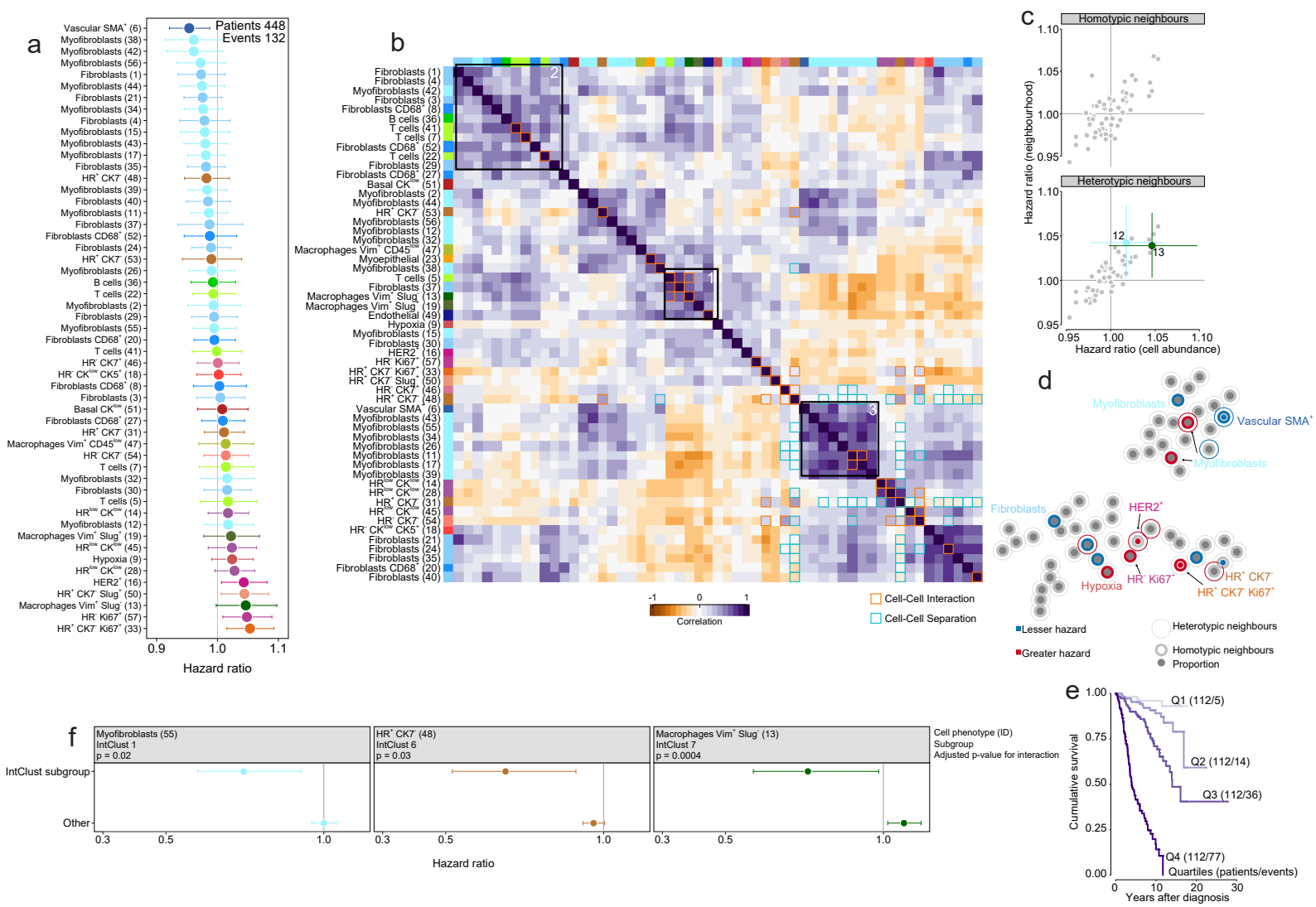

Macrophages Vim+ CD45low
Endothelial
Myoepithelial
Myofibroblasts
Vascular SMA+
HR- CKlow CK5+
Basal CKlow
Fibroblasts
HER2+
B cells
HRlow CKlow
HR+ CK7- Slug+
T cells
HR- CK7+
Fibroblasts CD68+
HR+ CK7-
Hypoxia
HR- Ki67+
HR- CK7-
HR+ CK7- Ki67+
Macrophages Vim+ Slug-
Macrophages Vim+ Slug+
tSNE 2
tSNE 1


ncells
1,000
20,000
60,000


IntClust 1
% ER+ 96.0
% HER2+ 24.0
Median survival:
% 5 year: 78.3
% 10 year: 64.9
n=25
IntClust 2
% ER+ 94.7
% HER2+ 10.5
Median survival:
% 5 year: 66.9
% 10 year: 60.2
n=19
IntClust 3
% ER+ 96.4
% HER2+ 5.4
Median survival:
% 5 year: 92.3
% 10 year: 85.2
n=56
IntClust 4+
% ER+ 100
% HER2+ 8.8
Median survival:
% 5 year: 98.0
% 10 year: 87.1
n=57
IntClust 4-
% ER+ 0
% HER2+ 51.9
Median survival:
% 5 year: 73.4
% 10 year: 68.2
n=27


IntClust 5
% ER+ 53.3
% HER2+ 100
Median survival:
% 5 year: 70.4
% 10 year: 60.0
n=30
IntClust 6
% ER+ 100
% HER2+ 4.5
Median survival:
% 5 year: 81.3
% 10 year: 74.5
n=22
IntClust 7
% ER+ 97.1
% HER2+ 0
Median survival:
% 5 year: 85.1
% 10 year: 80.6
n=35
IntClust 8
% ER+ 100
% HER2+ 4.2
Median survival:
% 5 year: 88.9
% 10 year: 78.2
n=48
IntClust 9
% ER+ 87.1
% HER2+ 9.7
Median survival:
% 5 year: 60.6
% 10 year: 60.6
n=31
IntClust 10
% ER+ 20.0
% HER2+ 17.5
Median survival:
% 5 year: 73.2
% 10 year: 70.3
n=40


Mutations
Amplifications
Deletions
Alteration frequency
250
200
150
100
50
0
HR+ CK7- (48)
HR+ CK7- (31)
HRlow CKlow (28)
Fibroblasts (21)
Fibroblasts (35)
Fibroblasts (24)
Myofibroblasts (11)
Myofibroblasts (17)
Myofibroblasts (34)
HR- CK7+ (46)
HR- CK7- (54)
HR+ CK7+ (53)
Myofibroblasts (38)
Myoepithelial (23)
Myofibroblasts (12)
Myofibroblasts (32)
Macrophages Vim+ CD45low (47)
HR- CKlow CK5+ (18)
Hypoxia (9)
Basal CKlow (51)
Fibroblasts (30)
HR- Ki67+ (57)
HER2+ (16)
Fibroblasts (37)
Macrophages Vim+ Slug- (13)
T cells (5)
Median proportion
CBFB
MAP3K1
PIK3CA
MAP2K4
GATA3
CTCF
MEN1
AHNAK2
THADA
MLL2
NOTCH1
COL6A3
RPGR
TP53
RB1
ATL2
KLF11
USP25
MRPS18A
BYSL
DNTTIP2
HEY1
INADL
PRDM1
SAMD8
BCL2
TNFRSF1A
B4GALNT3
CCNE1
CD274
MYB
IGF1R
PRKCI
PIK3CA
FBXL18
NUPL2
FKBP9
PLEKHA6
ARID4B
ZNF217
ERBB2
TRPS1
MYC
MRC2
TUBD1
ZNF652
CCND1
RSF1
ARID1A
ATM
APOBEC3B
BRCA1
CDKN2A
PBRM1
B2M
PTEN
Gene
ER-positive, median proportion
ER-negative, median proportion
Gain
Amplification
Non-synonymous mutation
Depleted cluster
Enriched cluster
Adjusted p-value < 0.05
Heterozygous deletion
Homozygous deletion


a
Patients 448
Events 132
Hazard ratio
b
Correlation
Cell-Cell Interaction
Cell-Cell Separation
c
Homotypic neighbours
Heterotypic neighbours
Hazard ratio (neighbourhood)
Hazard ratio (cell abundance)
d
Lesser hazard
Greater hazard
Heterotypic neighbours
Homotypic neighbours
Proportion
e
Q1 (112/5)
Q2 (112/14)
Q3 (112/36)
Q4 (112/77)
Quartiles (patients/events)
Cumulative survival
Years after diagnosis
f
Cell phenotype (ID)
Subgroup
Adjusted p-value for interaction
IntClust subgroup
Other
Hazard ratio