Molecular phenotyping of breast cancer cell lines at the single-cell level for automated cancer diagnosis and prediction of drug sensitivity

AUTHORS

Gambardella G[1,2,*], Viscido G[1,2,*], Tumaini B[1], Isacchi A[3], Bosotti R[3], di Bernardo D.

AFFILIATIONS

[1]Telethon Institute of Genetics and Medicine, Naples, Italy
[2]University of Naples Federico II, Department of Chemical Materials and Industrial Engineering, Naples, Italy
[3]NMS Oncology, Nerviano Medical Sciences, NMS Group, 20014, Nerviano, Milan, Italy

CORRESPONDENCE:

dibernardo@tigem.it

***These authors contributed equally to this work.**

ABSTRACT

Brest Cancer (BC) patient stratification is mainly driven by receptor status and histological grading and subtyping, with about twenty percent of patients for which absence of any actionable biomarkers results in no clear therapeutic intervention to apply. Here, we evaluated the potentiality of single-cell transcriptomics for automated diagnosis and drug treatment of breast cancer. We transcriptionally profiled 35,276 individual cells from 33 BC cell-lines covering all main BC subtypes to yield a Breast Cancer Single Cell Atlas. We show that single cell transcriptomics can successfully detect clinically relevant BC biomarkers and that atlas can be used to automatically predict cancer subtype and composition from a patient's tumour biopsy. We found that BC cell lines arbour a high degree of heterogeneity in the expression of clinically relevant BC biomarkers and that such heterogeneity enables cells with differential drug sensitivity to co-exist even within a genomically stable isogenic cell line. Finally, we developed a novel bioinformatics approach named DREEP (DRug Estimation from Expression Profiles) to automatically predict responses to more than 450 anticancer agents starting from single-cell transcriptional profiles. We validated DREEP both in-silico and in-vitro by selectively inhibiting the growth of the HER2-deficient subpopulation in the MDAMB361 cell line. Our work shows transcriptional heterogeneity is common, dynamic and plays a relevant role in determining drug sensitivity. Moreover, our Breast Cancer Single Cell Atlas and DREEP approach are a unique resource for the BC research community and to advance the use of single-cell sequencing in the clinics.

MAIN TEXT

Introduction

The main roadblock to make personalized medicine a clinical reality is the lack of standardized and easy-to-measure biomarkers to predict which drug or drug combination could benefit the patient starting from a tumour biopsy. Indeed, several multigene assays such as MammaPrint (*1*), Oncotype DX (*2*, *3*) and PAM50 (*4*) have been recently proposed to classify Breast Cancer (BC) tumour types and its risk of relapse (*5*). However, the clinical usefulness of these gene expression signatures is

principally limited to estrogen receptor positive breast cancers (*5*) and while their prognostic value is supported by different evidences (*6*), their ability in predicting drug sensitivity of the tumour is modest or null. As a consequence, in clinics BC patient stratification is still mainly driven by receptor status and histological grading and subtyping (*5*), with about twenty percent (*7*) of patients for which absence of any actionable biomarkers (*e.g.* triple negative breast cancer) results in no clear therapeutic intervention strategy. Moreover, even when a targeted treatment option is available, drug resistance may arise (*5*), partly because of rare drug tolerant subpopulations of cells characterized by different transcriptional or mutational states (*8–10*). Combination therapies have been long proposed to prevent drug resistance (*11*), but there is no general approach on how to choose the appropriate drugs.

Determining tumor transcriptional heterogeneity and its impact on drug response is essential to understand the mechanisms by which drug resistance arise and aid the development of more-effective personalized therapies. To date, genomic and non-genomic biomarkers of drug resistance have been found only for a restricted number of drugs and as a result, mechanisms of drug resistance remain poorly understood (*12–14*). Nevertheless, expression-based biomarkers measured from population-based RNA-sequencing (a.k.a. bulk RNA-seq) have been shown to be the most powerful predictors of drug response in vitro with far more predictive power than other –omics profiles (*13–15*). However, whole tumour tissue RNA sequencing averages out tumour heterogeneity and can miss rare cancer cell populations, which could instead confer drug resistance. On the contrary, single-cell transcriptomics yields a molecular profile of each individual cell and thus takes tumour heterogeneity into account (*16*, *17*). Moreover, it offers a rapid and comprehensive molecular phenotyping of the tumour at affordable costs (*16*, *17*), thus making it a prime candidate for routine clinical applications.

Here, we transcriptionally profiled 35,276 individual cells from 33 BC cell-lines covering all BC subtypes to assess transcriptional heterogeneity and to evaluate its impact on drug response. We organized single cell profiles in an atlas available online. We show that single cell transcriptomics can be successfully used in BC to capture the overall expression of clinically relevant markers and to automatically assess cancer subtype(s) from single cell sequencing of a patient's tumour biopsy. Moreover, we show that variability and plasticity in gene expression may enable cells with differential drug sensitivity to co-exist within the same cell-line. Finally we link results from large-scale in vitro drug screening (*13*, *14*) to our BC single cell data to develop a unique and innovative tool named DREEP (DRug Estimation from Expression Profiles) to automatically predict responses to more than 450 anticancer agents starting from single-cell transcriptional profiles.

RESULTS

1. Single-cell Transcriptome Profiling of Breast cancer cell-lines.

We performed single cell RNA-sequencing (scRNA-seq) of 32 breast cancer cell-lines (Figure 1A and Supplementary Table 01) and 1 basal-like normal breast epithelium cell-line (*18*) by means of the Drop-seq approach (*17*). Following pre-processing to filter out low quality cells and duplets (Methods and Supplementary Fig 02), we retained a total of 35,276 cells, with an average of 1,069 cells per cell-line and 3,248 genes captured per cell (Supplementary Fig. 01 and Supplementary Table 02).

We next generated an atlas encompassing all 33 cell-lines, as shown in Figure 1A and available online. This was achieved by combining data across cell-lines with the *gf-icf* pipeline (*19*), which performs count normalization, feature selection and dimensionality reduction (*20*) of the profiled cells. In the atlas, cell lines derived from the same cancer subtypes tend to cluster together, while being separated from the other subtypes (Figure

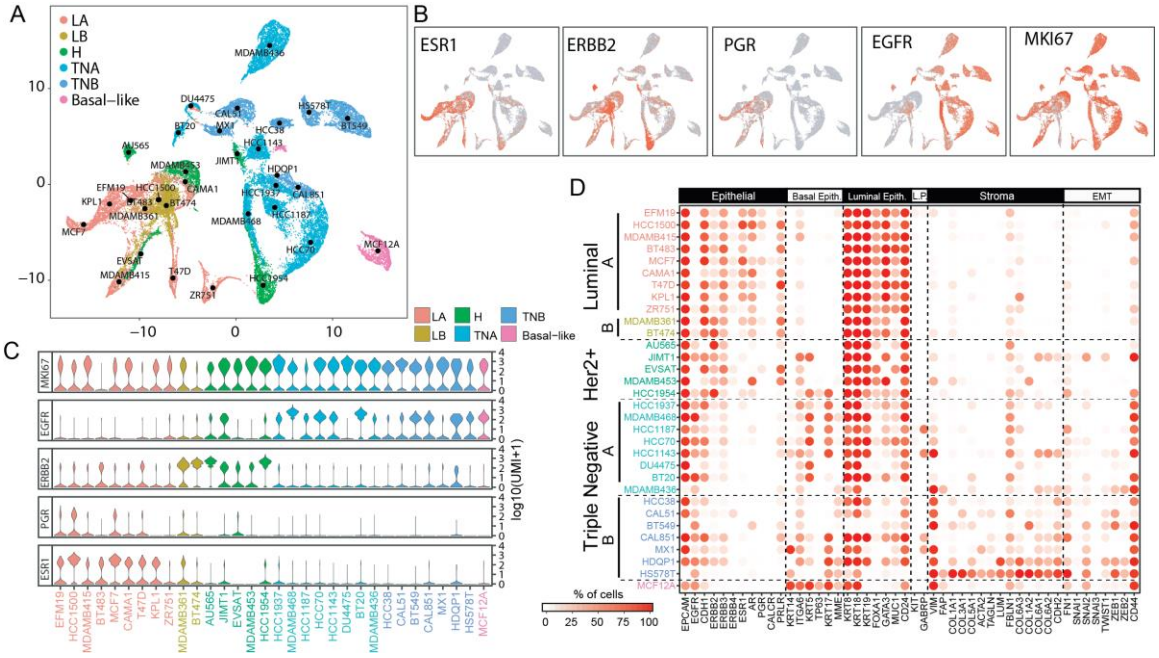

Figure 1 – The Breast Cancer Single Cell Atlas. (**A**) Graphical representation of single-cell transcriptomics of 33 cell lines for a total of 35,276 cells color-coded according to cancer subtype (LA=Luminal A, LB=Luminal B, H=Her2 positive, TNA = Triple Negative A, TNB = Triple Negative B). (**B**) Expression levels of the indicated biomarker genes in individual cells, with red indicating expression. (**C**) Expression distribution across BC CCLs of the same 5 BC marker in (C). (**D**) Dotplot of biomarker genes along the columns for each of the 33 sequenced cell-lines along the rows. Biomarker genes are grouped by type (Basal Epith. = Basal Epithelial, Luminal Epith. = Luminal Epithelial, L.P. = Luminal Progenitor, EMT = Epithelial to Mesenchymal Transformation).

1A): luminal BC cell lines form a big "island" with multiple "peninsulas" with intermixing of cells from distinct cell lines (Figure 1A,B); on the contrary, TNBC cells give rise to an "archipelago" with cells from the same cell-line grouped into distinct islands, thus suggesting that TNBC cell-lines represent instances of distinct diseases.

Single-cell expression of clinically relevant biomarkers including ESR1, PGR, ERBB2 (a.k.a. HER2) and EGFR across the different cell-lines is in agreement with their reported status, as measured by immunohistochemistry in the respective cell-lines (*18*) (Figure 1C). The majority of TNBC cells specifically expressed EGFR as expected but also exhibited higher expression of MKI67 (a.k.a. Ki-67), a clinical proliferation marker, than luminal breast cancer cells (*21*) (Figure 1B,C).

To gain further insights into each cancer cell-line, we analysed the expression of 48 additional literature-based stromal and epithelial biomarkers of clinical relevance (*22*), as reported in Figure 1D. Luminal cell-lines highly express luminal epithelium genes, but neither basal epithelial nor stromal markers; on the contrary, triple-negative BC cell-lines (11 out of 15) show a basal-like phenotype with the expression of at least one of keratin 5, 17 or 14 (*23*, *24*), with triple-negative subtype B (TNB) cell-lines also expressing VIM and COL6A1/2/3 genes (*18*). Interestingly, two out of five HER2+ cell-lines (JIMT1 and HCC1954) are both located in the atlas in the triple-negative "archipelago" and express keratin 5 (Figure 1A,D). Expression of keratin 5 in Her2-positive tumours has been shown to be linked to poor prognosis and trastuzumab resistance (*25*). Indeed, both cell have either no or very low response to anti-HER2 treatments, including Trastuzumab and Laptinib (*26*). Finally, as already reported by immunohistochemical analysis (*27*), the non-tumorigenic MCF12A cell-line lacks expression of ESR1, PGR and HER2 and displays a

basal-like phenotype characterized by the expression of all basal-like marker genes including keratin 5, 14, 17 and TP63.

Overall, these results show that single cell transcriptomics can be successfully used to capture the overall expression of clinically relevant markers in breast CCLs, including the status of receptors relevant for BC patient care and treatment, whose status is usually

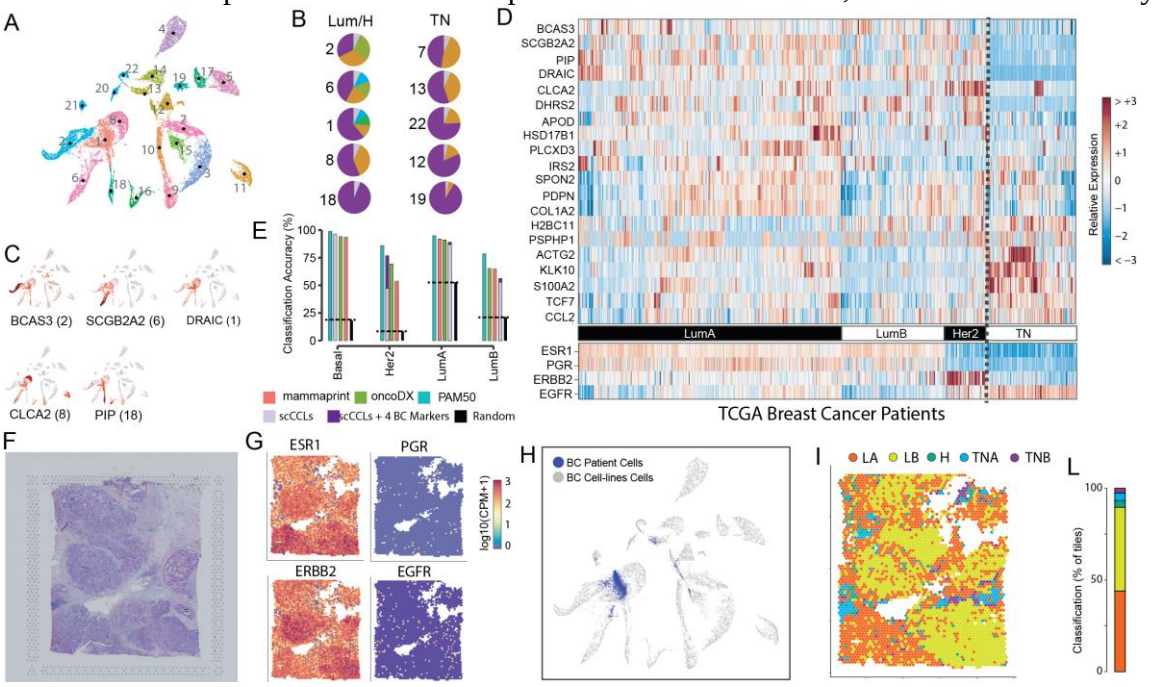

Figure 2 –Automatic classification of patients' tumour cells. (**A**) Graphical representation of 35,276 cells color-coded according to their cluster of origin. Clusters are number from 1 to 22. (**B**) For the indicated cluster, the corresponding pie-chart represents the cluster composition in terms of cell lines. Cell lines in the same pie-chart are distinguished by colour. Only the top 10 most heterogenous clusters are shown. Cluster 2 is the most heterogeneous while cluster 19 is the most homogeneous. (**C**) Expression levels in the atlas of the five luminal biomarkers identified as the most differentially expressed in each of the five luminal clusters (1, 2, 6, 8 and 18). (**D**) Expression of the atlas-derived biomarkers in the biopsies of 937 breast cancer patient from TCGA. (**E**) Accuracy in classifying tumour subtype for 937 patients from TCGA by using either PAM50, oncoDX, mammaprint, or the 22 atlas derived biomarker genes (scCCL) alone or augmented with HER2, ESR1, PGR and EGFR genes (scCCL + 4 BC markers). (**F**) Tissue-slide of tumour biopsie sequenced by 10x spatial transcriptomics. (**G**) Spatial expression of ESR1, PGR, ERBB2 and EGFR genesn. (**H**) Pseudo-cells sequenced with spatial transcriptomic are embedded in the BC atlas to predict their tumour type using K-nn algorithm. (**I**) Classification of each pseudo cell to show predicted tumour type in the spatial context. (**L**) Quantification of (I) where the percentage of pseudo cells predicted for each tumour type is reported.

assessed by immunohistochemistry assay.

2. The BC single-cell atlas identifies clinically relevant transcriptional signatures.

Clustering of the 35,276 single-cells in the atlas identified 22 clusters, as shown in Figure 2A. Within the island composed by luminal BC cells, cells did not cluster according to their cell-line of origin, indeed four out of the five luminal clusters contain cells from distinct cell-lines (Figure 2B and Supplementary Figure 02). On the contrary, triple-

negative cancer cell lines clustered according to their cell-line of origin, with each cluster containing mostly cells from the same cell-line, hinting at quite distinct diseases.

We identified the gene most specifically expressed among the cells in the same cluster for a total of 22 biomarkers, one for each cluster (Figure 2C and Supplementary Table 03 and Supplementary Figure 03). Literature mining confirmed prevalence and clinical relevance of the markers thus identified (Supplementary Table 03): for example, clusters in the luminal island (Figure 2C) were associated to genes involved in cancer progression (BCAS3 (*28*, *29*) cluster 2), dissemination (SCGB2A2 (*30*, *31*) cluster 6), proliferation (DRAIC (*32*, *33*) cluster 1), migration and invasion (CLCA2 (*34*, *35*) cluster 8 and PIP (*36*) cluster 18). Interestingly, whereas DRAIC is correlated with poorer survival of luminal BC patients (*33*), both CALCA2 and PIP are significantly associated with a favourable prognosis (*34*, *35*, *37*). Indeed, CLCA2 was shown to be downregulated in several primary breast tumours, and loss of CLCA2 was associated with tumorigenicity and invasion potential (*35*, *38*) while overexpression showed decreased proliferative, migrating and invasive features (*34*). To confirm the role of CALC2 and PIP in a favourable prognosis, we performed a survival analysis (*39*) on 503 luminal BC patients (e.g. ESR1 positive) which transcriptional profiles were measured before any drug treatment and confirmed an increased survival time for patient highly expressing these two genes (Supplementary Figure 04).

To further examine the clinical relevance of these 22 biomarkers, we analysed their expression across 937 breast cancer patients from the TGCA collection encompassing all four BC types. As shown in Figure 2D, there is a clear difference in their expression among Luminal A, Luminal B, Her2 positive and Triple Negative patients. Moreover, it is possible to distinguish subtypes within each category, which may lead to novel diagnostic biomarkers (Figure 2D and Supplementary Figure 05). For example, one subset of triple-negative patients strongly expresses KLK10 that has already been associated with poor prognosis, poor response to tamoxifen treatment (*40*) and identified as potential target to reverse trastuzumab resistance (*41*). Whereas a second subset of triple-negative patients is characterised by ACTG2 expression, which overexpression has been already linked in BC cells to proliferation (*42*) and platinum-based chemotherapy sensitivity (*43*, *44*) including paclitaxel (*45*, *46*). Finally, to prove the prognostic value of our 22 biomarker genes we compared their performances in classifying BC type from bulk RNA-seq data (see methods) against MammaPrint (*1*), Oncotype DX (*2*, *3*) and PAM50 (*4*) gene signatures (Figure 2E).

Altogether, these analyses confirm that the single cell BC cell-line atlas allows identifying clinically relevant transcriptional signatures for BC patient useful for patient stratification and tumour type classification.

3. The BC atlas as a reference for automated cancer diagnosis

We hypothesised that the BC atlas could be used as a reference against which to compare single cell transcriptomics data from a patient's tissue biopsy to perform automatic diagnosis of cancer and assessment of heterogenity. We thus developed an algorithm for automatic cancer type classification. The algorithm first projects the cancer transcriptional profile onto the BC atlas as if it were a new "cell" and then assigns it to a cancer subtype according to the one most common among its neighbouring cells (Methods). We first tested the ability of the algorithm in classifying the very cells in the atlas starting from their

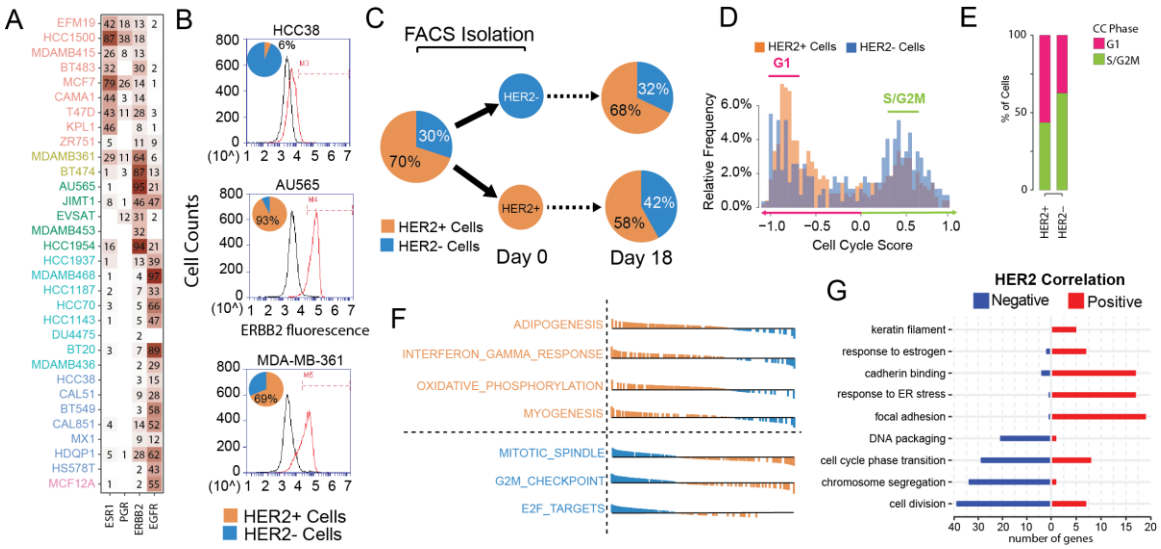

Figure 3 – Gene expression heterogeneity is isogenic breast cancer cell lines. (**A**) Percentage of cells expressing the indicated genes in each of the 33 cell lines we sequenced. (**B**) FACS analysis of HCC38, MDAMB361 and AU565 cell-lines stained with antibody against Her2. (**C**) Expression of HER2 protein in MDAMB361 cells is dynamic and re-established in about 3 weeks. (**D**) Predicted cell cycle phase for both HER2+ and HER2- cell subpopulation of MDAMB361. (**E**) Quantification of (D) showing the percentage of HER2+ or HER2- MDAMB361 cells predicted to be in either G1 or s/G2M phase. (**F**) Gene set enrichment analysis performed against the ranked list of differentially expressed genes obtained by comparing the single-cell transcriptional profiles of HER2+ versus HER2- MDAMB361 cells. (**G**) Gene ontology enrichment analysis of genes whose expression is either positively or negatively correlated with HER2 in MDAMB361 cell line.

transcriptional profiles by using a standard cross-validation strategy where 75% of the BC atlas' cells make up the training set and the remaining 25% represent the test set. On this dataset, the algorithm correctly assigned the cancer subtype of the cells 92% of the times (Supplementary Figure 06).

We next tested the algorithm on a tissue biopsy of a breast cancer patient. Specifically, we downloaded from the 10X Genomics public repository, the spatial transcriptomic data from a tissue biopsy of a patient diagnosed $ESR1^{+}/ERBB2^{+}$ ductal carcinoma (Figure 2F,G). The dataset consists of 3,615 transcriptional profiles, each obtained from a different tissue "tile" of size 100um x 100um x 100 um (Figure 2F,G). The algorithm projected each of the 3,615 profiles into the cell-line BC atlas, as shown in Figure 2H, and then assigned a tumour type to each tile, as shown in Figure 2I,L. Interestingly, although 89% of the tiles were correctly classified as luminal A or B, the remaining 11% was classified either as HER2 positive or Triple Negative, which could be important for choosing the right therapy and to predict possible drug resistance. The same results were confirmed in a second tissue biopsy of the same patient (Supplementary Figure 07).

4. Clinically relevant biomarkers exhibit heterogenous and dynamic expression in BC cell lines.

Single-cell transcriptomics enables measurement of transcriptional heterogeneity present in a tissue or cell population. Indeed, one of the motivating factors of this study was to assess whether breast-cancer cell lines exhibit transcriptional heterogeneity, despite being genomically stable, and if so, what is the impact on drug response. We found that clinically

relevant receptors are indeed heterogeneously expressed across cells belonging to the same cell lines, as assessed by computing the percentage of cells expressing each receptor (*i.e.* > 1 count), and as reported in Figure 3A, despite a clear difference in their expression across cancer subtypes. Consider the seven Luminal B and HER2^{+} cell-lines present in the BC atlas, which by definition overexpress HER2: whereas more than 90% of cells in AU565, BT574 and HCC1954 cell-lines expresses *ERBB2*, the gen coding for the HER2 protein, in the remaining 4 cell-lines ERBB2 expression ranged from 31% (EVSAT) to 64% (MDAMB361). Interestingly, AU565, BT474, HCC1954, JIMT1 and MDAMB361 all have a gain in genomic copy number variation (CNV) of the locus containing the *ERBB* (*47*), nevertheless only 64% of MDAMB361 and 46% of JIM1 cells express it, according to single-cell data. To exclude the possibility that these results were artifacts of single-cell RNA-sequencing technology, we first excluded variability in the depth of coverage across cell-lines as a possible confounding factor (Supplementary Figure 08). We then assessed HER2 protein levels by flow cytometry with an anti-HER2 antibody in AU565, MDAMB361 and HCC38 cell-lines. As shown in Figure 3B, single-cell transcriptional data are in agreement with the cytometric analysis, confirming the reliability of the former in measuring clinically relevant biomarkers. Interestingly, the MDAMB361 cell-line has a quite balanced expression of HER2 ranging between 64% (single-cell) to 69% (cytometry), however, the origin of this heterogeneity is difficult to explain as these cells are genomically stable (Supplementary Figure 09). To exclude hereditable genetic differences as a source of heterogeneity, we sorted MDAMB361 cells into HER2+ and HER2- subpopulation (see methods) and checked whether these homogenous subpopulations were stable or rather spontaneously give rise to heterogeneous population over time. As shown in Figure 3C, after three weeks in culture, both subpopulations re-established the original heterogeneity, demonstrating that HER2 expression in these cells is dynamic and driven by a yet undiscovered mechanism.

Recently, HER2+ circulating tumour cells (CTCs) isolated from a ER+/HER2− breast cancer patient were shown to spontaneously interconvert from HER2+ and HER2−, with cells of one phenotype producing daughters of the opposite one (*48*). To check for a role of cell-cycle in explaining heterogeneity in the MDAMB361 cell population, we bioinformatically predicted (see methods) from single cell transcriptomic data the cell cycle phase (*49*) of each cell in both the HER2+ and HER2- subpopulations. A higher proportion of HER2- cells was predicted to be in S /G2/M phases (Figure 3D) when compared to HER2+ cells (Figure 3E).This result is consistent with previous observations which reports cell cycle arrest in the G2/M phase following HER2 inhibition (*50*), suggesting that differences in cell-cycle are consequence rather than a cause of the heterogeneous HER2 expression.

We next set to identify additional biological processes differing between the two subpopulations. To this end, we first identified differentially expressed genes (DEGs) by comparing the single-cell transcriptional profiles of HER2+ cells against HER2- cells (Supplementary Table 04) and then performed Gene Set Enrichment Analyses (GSEA) (*51*) against the ranked list of DEGs. This analysis revealed seven significantly enriched pathways (FDR<10%) (Figure 3F and Methods): four pathways were upregulated in HER2+ cells and downregulated in HER2- cells and include adipogenesis, myogenesis and OXPHOS all indicative of EMT engagement in HER2+ cells (*52–54*); the remaining three out the seven pathways were related to cell-cycle and upregulated in HER- cells but

downregulated in HER2+ cells, consistently with our previous analysis indicating a difference in cell cycle phases among the two subpopulations. Overexpression of HER2 has already been linked to increased EMT and metastasis in vitro and in vivo (*53*, *55*), however the dynamic nature of HER2+ has not been well explored because of the lack of cellular models. Here we identified the MDAM361 as a potential model to study HER2+ dynamics.

To further validate these findings, we performed a gene expression correlation analysis to identify the genes whose expression is significantly correlated or anti-correlated with HER2 across single cells of the MDMA361 cell lines (Figure 3G). 225 genes were positively correlated with HER2 expression (FDR<0.05) and Gene Ontology Enrichement Analysis of these genes revealed as highly significant (FDR<0.0001) terms related to cell-cell adhesion, in line with the observed EMT engagement in HER2 expressing cells. On the contrary, the 99 genes significantly anti-correlated with HER2 (FDR<0.05) were enriched for terms related to cell-cycle, again in agreement with alteration of cell cycle pathways in cells not expressing HER2. As expected HER2- showed an increased expression of genes related to DNA damage (*48*) while HER2+ cell programs were mostly related to cell migration and invasion.

5. Heterogeneity in gene expression affects the drug response of the cell population.

To investigate the role of heterogeneity in gene expression on drug response, we linked results from large-scale in vitro drug screening (*13*, *14*) with our single-cell data set (Figure 4A,B). To this end, we collected large-scale in vitro drug screening data (*13*, *14*) reporting the effect for 450 drugs on 658 cancer cell-lines from solid tumours, including the BC cell-lines profiled in this study. As show in Figure 4A, sensitivity of the BC cell-lines to HER2 inhibitors was highly correlated to the percentage of cells in the cell-line expressing *ERBB2* (*i.e.* the lower the proportion of cells expressing *ERBB2*, the lower the effect of the anti-HER2 drug). Interestingly, as reported in Supplementary Figure 10, we observed that receptor expression is substantially the same across cells expressing it, irrespective of the cell-line they belong to, except for cell-lines with CNVs including the gene of interest. This means that when assessing gene expression using “bulk” measurements (as opposed to single cell), the level of expression of a clinically relevant receptor does not imply that all of the cells express that receptor at the same measured “average” level, rather the measured bulk expression is proportional to the percentage of cells expressing it in the population. Furthermore, we found that the correlation between drug target expression and drug sensitivity holds true also for several targets (Figure 4B and Supplementary Table 04), thus confirming that variability in gene expression within cells of the same tumour may cause some cells to respond poorly to the drug treatment (Figure 4B).

Based on this observation, we developed DREEP (DRug Estimation from Expression Profiles), a novel bioinformatics tool that starting from single-cell transcriptional profiles allows to predict drug response at the single cell level. To this end, we first set to automatically detect expression-based biomarkers of drug sensitivity for over 400 drugs (*13*), including both targeted and chemotherapeutic compounds, as schematised in Figure 4C,D (see Methods). Briefly, we crossed data from the Cancer Cell Line Encyclopedia (CCLE) on the response to 450 drugs across 658 cancer cell-lines from solid tumours with their gene expression profiles from bulk RNA-seq (see Methods). In the

CCLE, drug potency for each cell line is evaluated as the inverse of the Area Under the Curve (AUC) of the dose-response graph, with low values of the AUC indicating drug sensitivity, while high values implying drug resistance (Figure 3E). For each gene and for each drug, we computed the correlation between the expression of the gene across the 658 cell-lines with the drug potency in the same cell-lines. Hence, genes positively correlated with the AUC are potential markers of resistance (the more expressed the gene, the higher the concentration needed to inhibit growth), vice-versa, negatively correlated genes are markers of sensitivity (Figure 4C). In this way, we generated a ranked list of expression-based biomarkers of drug sensitivity and resistance for each of the 450 drugs (Figure 4C). We then used these biomarkers to predict drug sensitivity at the single-cell level (Figure 4D). To this end, for each cell in the single-cell BC atlas, the 250 genes most expressed in that cell were compared against the ranked list of biomarkers for each one of 450 drugs by means of Gene Set Enrichment Analysis (GSEA) (*51*). A negative enrichment score implies that highly expressed genes in the cell are enriched among biomarkers associated to sensitivity (Figure 4D), whereas a positive enrichment score implies enrichment of resistance-associated biomarkers. At the end of this process, each cell in the atlas is associated to the drug it is most sensitive to, or to no drug, if no significant enrichment score from GSEA is found (Figure 4D).

In order to assess the predictive ability of this approach, we applied it to the single-cell BC atlas and estimated its performance by checking how well we could predict sensitivity of the 32 BC cell lines to 86 drugs for which this information was publicly available from GDSC (Genomics of Drug Sensitivity in Cancer) (*56*). Hence, we assigned to each of the 2,752 cell-line/drug combinations (32 cell-lines x 86 drugs), the value 1 if the cell-line is sensitive to the drug according to the experimental data (see methods), or 0 otherwise. We then used our algorithm to predict drug sensitivity for each combination and compared our predictions to the experimental data. To convert single-cell predictions to bulk predictions for the cell-line, we simply used the percentage of cells in the cell-line deemed to be sensitive to the drug by the algorithm. The results are shown in Figure 4E with the percentage of correctly predicted cell-line/drug combinations (y-axis) as a function of the percentage of cells predicted to be sensitive to the drug in the cell-line (x-axis). It can be observed that when the method predicts that more than 95% of the cells as sensitive to the drug, the prediction is always correct, and in any case always better than random (dashed line). As a further validation, we also computed for each of the 86 drugs the correlation between the measured IC50 values from GDSC (*56*) and the percentage of cells predicted to be either resistant, sensitive or not classified to by our method (Figure 4F).

To experimentally validate our algorithm, we turned to the MDAMB361 cell-line for which we found coexistence of two distinct and dynamic HER2+ and HER2− cell subpopulations. We applied our algorithm to each subpopulation to identify drugs able to selectively inhibit growth of the HER2- subpopulation but not the HER2+ one. As shown in Supplementary Figure 11, 46 drugs (FDR < 1%) were predicted to specifically inhibit HER2- cells but not HER2+ cells, whereas 152 (FDR < 5%) drugs were predicted to inhibit both subpopulations. Counterintuitively, anti-HER2 drugs such as afatinib and lapatinib were among the latter and deemed to be effective on both subpopulations, despite the HER2- not expressing the target. Among the HER2- specific drugs, there were few targeted drugs and several chemotherapeutic agents such as etoposide. To check the validity of our

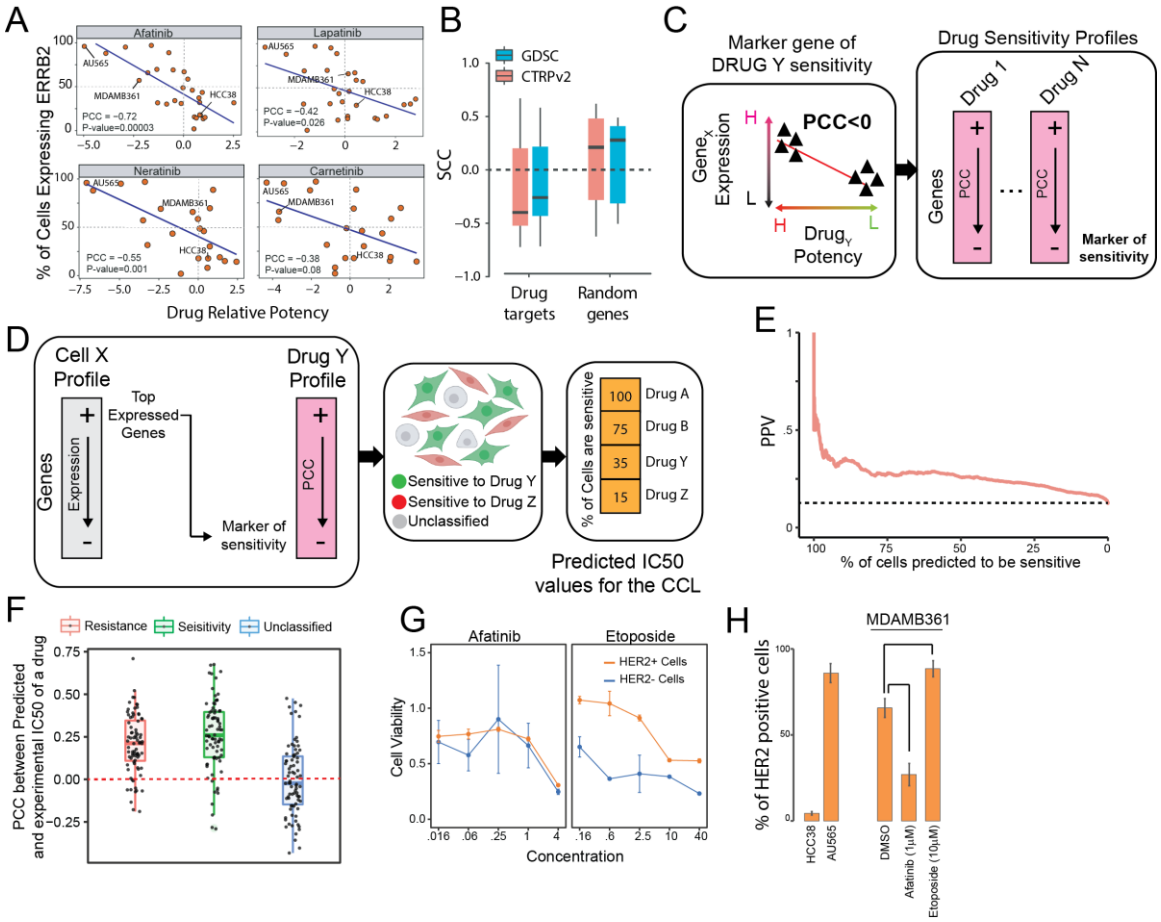

Figure 4 – Prediction of drug treatment from single-cell gene expression profiles. (**A**) Relationship between percentage of cells expressing ERBB2 in the cell-line [y-axis] and the drug potency in the same cell line [x-axis] for the indicated molecules. The more negative the value on the x-axis, the more potent the drug. Drug potency data were retrieved from the CTRPv2 or GDSC databases. Each dot represents a cell-line we have sequenced with single cell transcriptomics. (**B**) The negative correlation between percentage of cells expressing the drug target and the potency of the drug holds for most of the targeted drugs across cell lines. Box-plot of Spearman correlation coefficient (SCC) for two different drug potency databases when correlating against cognate drug targets or random genes . (**C**) Construction of the ranked list of drug sensitivity biomarkers for 450 drugs. For each gene and for each drug, the expression of the gene is correlated with the potency of a drug expressed as a function of Area Under the Curve (AUC) across 658 cell-lines. (**D**) The top 250 most expressed genes of a cell are used as input for a Gene Set Enrichment Analysis (GSEA) against the ranked list of biomarkers for each one of 450 drugs, to predict single cell drug sensitivity. At the end of the process, each cell in the sample is associated to the drug it is most sensitive to, or to no drug, if no significant enrichment score from GSEA is found. Finally, for each of the 450 drugs, the number of cells predicted to be either sensitive, resistant or not classified in the considered sample are estimated. (**E**) Validation of the computational method in (D) using the Breast Cancer Single Cell atlas data to predict drug sensitivity to 86 drugs for which the corresponding half maximal inhibitory concentration (IC50) was available from the GDSC database. The PPV (Positive Predicted Value) is shown as a function of the percentage of cells predicted to be sensitive for each drug-BC CCL interaction. Dashed line represents the performance of a random algorithm. (**F**) Box-plot of the Pearson correlation coefficient (PCC) between the predicted and measured IC50. (**G**) Dose-response curve against for afatinib and etoposide on sorted MDAMB361 cell populations (triplicate experiment). (**H**) Percentage of HER2+ cells in MDAMB361 after treatment with either afatinib or etoposide. HCC38 and AU565 were used as negative and positive controls, respectively. The experiment was performed in triplicate.

338

predictions, we selected afatinib and etoposide for experimental validation. MDAM361 cells were first sorted by FACS into HER2+and HER2- cells and then dose-response curves for both drugs were performed on the two subpopulations, as shown in Figure 4G (see Methods). As predicted, HER2- cells were more sensitive to etoposide than HER+ cells, while Afatinib was effective on both subpopulations. In agreement with this surprising observation, Jordan et al (*48*) found that circulating tumour cells from a patient sorted into HER2- and HER2+ subpopulations were equally sensitive to Lapatinib (another HER2 inhibitor), but no mechanism was put forward. This apparent contradiction may be explained by dynamic interconversion of HER2+ to HER2- cells, indeed afatinib acts only on HER2+ cells, however since HER2- cells will interconvert to HER2+ cells during the treatment, then afatinib will apparently work also on HER2- cells.

To further validate this finding, we treated the MDAM361 cell-lines (without sorting) with afatinib and etopiside and then counted by cytofluorimentry the percentage of HER2+ and HER2- cells before and after the treatment. As shown in Figure 4H, etoposide increased the percentage of HER2+ cells, in agreement with the increased sensitivity of HER2- cells to this treatment. Whereas, afatinib decreased the percentage of HER2+ cells, confirming that its effect is on HER2+ cells only.

All together these results show that our algorithm can predict drug sensitivity from single-cell transcriptional profiles and that heterogeneity in gene expression does play a significant role in how the cell population will respond to the drug treatment.

Discussion

In this study we provide the first transcriptional characterization at single cell level of a panel of 33 breast cell-lines, including all main BC subtypes and one basal-like breast cell-line. We show that although limited in depth and breadth of coverage, single cell transcriptomics can be used to capture the overall expression of clinically relevant markers in breast CCLs, including the status of receptors relevant for BC patient care usually assessed by immunohistochemistry assay in clinics. Differently from bulk transcriptomics, single-cell transcriptomics has the advantage to enable the quantification of the transcriptional heterogeneity present in a population of cells. Indeed, we show that despite being genomically stable, breast-cancer cell lines are transcriptional heterogeneous with cells within the same cell-line displaying a marked degree of heterogeneity in the expression of clinically relevant BC receptors including HER2, ESR1, PGR and EGFR. Accordingly, we observed dynamic plasticity in the regulation of HER2 expression in the MDAMB361 cell line, a phenomenon recently observed also in circulating tumour cells of a BC patient (*48*) with striking consequences on drug response. Although the mechanisms by which those cells are able to spontaneously interconvert from HER2+ to HER2− and vice versa remain yet undiscovered, here we found that HER2 deficient cells show the activation of several transcriptional program related to DNA damage (*48*) and cell cycle, while HER2+ cells were mostly characterized by the expression of genes related to cell migration and invasion.

Moreover, by linking results from large-scale in vitro drug screening (*13*, *14*) with our breast single-cell dataset, we show that variability in gene expression of breast cancer receptors within cells of the same cell-line may cause some cells to respond poorly to the drug treatment

Based on this observation, we developed DREEP that uses single cell transcriptomics to predict the most effective drug therapy to apply at single cell resolution. We show DREEP was able to correctly identify etoposide to selectively inhibit the growth of the HER2- subpopulation of MDAMB361 cell line.

Thanks to single-cell transcriptomics, a deeper characterization of the transcriptional heterogeneity can now be modelled in different clinical contexts. However, the ability to associate this heterogeneity to drug response is a prerequisite for a deeper understanding of the mechanisms that govern drug resistance. Here we described the landscape of cellular diversity across 33 cell lines of breast cancer, highlighting how transcriptional heterogeneity can affect drug response and showing how exploit it from single cell transcriptomics to predict drug response. However, further studies of these mechanisms are needed to provide a better understanding of intra-tumour heterogeneity and transform this understanding into novel treatment strategies to avoid drug resistance.

References

1. F. Cardoso *et al.*, 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *N. Engl. J. Med.* **375**, 717–729 (2016).

2. J. A. Sparano *et al.*, Prospective Validation of a 21-Gene Expression Assay in Breast Cancer. *N. Engl. J. Med.* **373**, 2005–2014 (2015).

3. J. A. Sparano *et al.*, Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. *N. Engl. J. Med.* **379**, 111–121 (2018).

4. M. C. U. Cheang *et al.*, Defining Breast Cancer Intrinsic Subtypes by Quantitative Receptor Expression. *Oncologist*. **20**, 474–482 (2015).

5. N. Harbeck *et al.*, *Breast cancer* (2019), vol. 5.

6. F. Andre *et al.*, Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: ASCO Clinical Practice Guideline Update—Integration of Results From TAILORx. *J. Clin. Oncol.* **37**, 1956–1964 (2019).

7. W. D. Foulkes, I. E. Smith, J. S. Reis-Filho, Triple-Negative Breast Cancer. *N. Engl. J. Med.* **363**, 1938–1948 (2010).

8. S. V. Sharma *et al.*, A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations. *Cell*. **141**, 69–80 (2010).

9. S. M. Shaffer *et al.*, Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*. **546**, 431–435 (2017).

10. S. Ebinger *et al.*, Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia. *Cancer Cell*. **30**, 849–862 (2016).

11. J. R. Pritchard *et al.*, Defining principles of combination drug mechanisms of action. *Proc. Natl. Acad. Sci.* **110**, E170–E179 (2013).

12. M. J. Garnett *et al.*, Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*. **483**, 570–575 (2012).

13. M. G. Rees *et al.*, Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol*. **12**, 109–116 (2016).

14. F. Iorio *et al.*, A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*. **166**, 740–754 (2016).

15. J. C. Costello *et al.*, A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol*. **32**, 1202–1212 (2014).

16. G. X. Y. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 1–12 (2017).

17. E. Z. Macosko *et al.*, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. **161**, 1202–1214 (2015).

18. X. Dai, H. Cheng, Z. Bai, J. Li, Breast cancer cell line classification and Its relevance with breast tumor subtyping. *J. Cancer*. **8**, 3131–3141 (2017).

19. G. Gambardella, D. di Bernardo, A Tool for Visualization and Analysis of Single-Cell RNA-Seq Data Based on Text Mining . *Front. Genet.* . **10** (2019), p. 734.

20. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (2018) (available at http://arxiv.org/abs/1802.03426).

21. N. A. Soliman, S. M. Yussif, Ki-67 as a prognostic marker according to breast cancer molecular subtype. *Cancer Biol. Med.* **13**, 496–504 (2016).

22. M. Karaayvaz *et al.*, Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat. Commun.* **9** (2018), doi:10.1038/s41467-018-06052-0.

23. S. Badve *et al.*, Basal-like and triple-negative breast cancers: A critical review with an emphasis on the implications for pathologists and oncologists. *Mod. Pathol.* **24**, 157–167 (2011).

24. B. Gusterson, Do “basal-like” breast cancers really exist? *Nat. Rev. Cancer*. **9**, 128–134 (2009).

25. B. Martin-Castillo *et al.*, Cytokeratin 5/6 fingerprinting in HER2-positive tumors identifies a poor prognosis and trastuzumab-resistant Basal-HER2 subtype of breast cancer. *Oncotarget*. **6**, 7104–7122 (2015).

26. S. Jernström *et al.*, Drug-screening and genomic analyses of HER2-positive breast cancer cell lines reveal predictors for treatment response. *Breast Cancer Targets Ther.* **9**, 185–198 (2017).

27. M. F. Sweeney, C. Sonnenschein, A. M. Soto, Characterization of MCF-12A cell phenotype, response to estrogens, and growth in 3D. *Cancer Cell Int.* **18**, 1–12 (2018).

28. A. E. Gururaj *et al.*, MTA1, a transcriptional activator of breast cancer amplified sequence 3. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 6670–6675 (2006).

29. M. Bärlund *et al.*, Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. *Genes Chromosom. Cancer*. **35**, 311–317 (2002).

30. B. K. Zehentner, D. Carter, Mammaglobin: A candidate diagnostic marker for breast cancer. *Clin. Biochem.* **37**, 249–257 (2004).

31. F. S. Al Joudi, Human mammaglobin in breast cancer: A brief review of its clinical utility. *Indian J. Med. Res.* **139**, 675–685 (2014).

32. M. Sun, S. S. Gadad, D. S. Kim, W. L. Kraus, Discovery, Annotation, and Functional Analysis of Long Noncoding RNAs Controlling Cell-Cycle Gene Expression and Proliferation in Breast Cancer Cells. *Mol. Cell*. **59**, 698–711 (2015).

33. D. Zhao, J. T. Dong, Upregulation of long non-coding RNA DRAIC correlates with adverse features of breast cancer. *Non-coding RNA*. **4**, 1–9 (2018).

34. Y. Y. Qiang *et al.*, Along with its favorable prognostic role, CLCA2 inhibits growth and metastasis of nasopharyngeal carcinoma cells via inhibition of FAK/ERK signaling. *J. Exp. Clin. Cancer Res.* **37**, 1–14 (2018).

35. X. Li, J. K. Cowell, K. Sossey-Alaoui, CLCA2 tumour suppressor gene in 1p31 is epigenetically regulated in breast cancer. *Oncogene*. **23**, 1474–1480 (2004).

36. A. Urbaniak, K. Jablonska, M. Podhorska-Okolow, M. Ugorski,

P. Dziegiel, Prolactin-induced protein (PIP)-characterization and role in breast cancer progression. *Am. J. Cancer Res.* **8**, 2150–2164 (2018).

37. M. A. Debily *et al.*, A functional and regulatory network associated with PIP expression in human breast cancer. *PLoS One*. **4** (2009), doi:10.1371/journal.pone.0004696.

38. A. D. Gruber, B. U. Pauli, *Cancer Res.*, in press (available at http://cancerres.aacrjournals.org/content/59/21/5488.abstract).

39. B. Györffy *et al.*, An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res. Treat.* **123**, 725–731 (2010).

40. Z. Wang *et al.*, Identification of KLK10 as a therapeutic target to reverse trastuzumab resistance in breast cancer. *Oncotarget*. **7**, 79494–79502 (2016).

41. L.-Y. Luo, E. P. Diamandis, M. P. Look, A. P. Soosaipillai, J. A. Foekens, Higher expression of human kallikrein 10 in breast cancer tissue predicts tamoxifen resistance. *Br. J. Cancer*. **86**, 1790–1796 (2002).

42. V. Dugina, G. Shagieva, N. Khromova, P. Kopnin, Divergent impact of actin isoforms on cell cycle regulation. *Cell Cycle*. **17**, 2610–2621 (2018).

43. X. Lu *et al.*, Establishment of a Predictive Genetic Model for Estimating Chemotherapy Sensitivity of Colorectal Cancer with Synchronous Liver Metastasis. *Cancer Biother. Radiopharm.* **28**, 552–558 (2013).

44. K. Edfeldt, P. Hellman, G. Westin, P. Stalberg, A plausible role for actin gamma smooth muscle 2 (ACTG2) in small intestinal neuroendocrine tumorigenesis. *BMC Endocr. Disord.* **16**, 19 (2016).

45. C.-Z. Xu *et al.*, Gene and microRNA expression reveals sensitivity to paclitaxel in laryngeal cancer cell line. *Int. J. Clin. Exp. Pathol.* **6**, 1351–1361 (2013).

46. N. M. Verrills *et al.*, Alterations in γ-Actin and Tubulin-Targeted Drug Resistance in Childhood Leukemia. *JNCI J. Natl. Cancer Inst.* **98**, 1363–1374 (2006).

47. M. Ghandi *et al.*, Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*. **569**, 503–508 (2019).

48. N. V. Jordan *et al.*, HER2 expression identifies dynamic functional states within circulating breast cancer cells. *Nature*. **537**, 102–106 (2016).

49. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411 (2018).

50. Y. Yan *et al.*, A novel function of HER2/Neu in the activation of G2/M checkpoint in response to γ-irradiation. *Oncogene*. **34**, 2215–2226 (2015).

51. A. Subramanian *et al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–50 (2005).

52. D. Ishay-Ronen *et al.*, Gain Fat—Lose Metastasis: Converting Invasive Breast Cancer Cells into Adipocytes Inhibits Cancer Metastasis. *Cancer Cell*. **35**, 17-32.e6 (2019).

53. S. Ingthorsson *et al.*, HER2 induced EMT and tumorigenicity in breast epithelial progenitor cells is inhibited by coexpression of EGFR. *Oncogene*. **35**, 4244–4255 (2016).

54. C. D. Savci-Heijink *et al.*, Epithelial-to-mesenchymal transition status of primary breast carcinomas and its correlation with metastatic behavior. *Breast Cancer Res. Treat.* **174**, 649–659 (2019).

55. P. Gupta, S. K. Srivastava, HER2 mediated de novo production of TGFβ leads to SNAIL driven epithelial-to-mesenchymal transition and metastasis of breast cancer. *Mol. Oncol.* **8**, 1532–1547 (2014).

56. W. Yang *et al.*, Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–D961 (2013).

57. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2012).

58. J. Harrow *et al.*, GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* . **22**, 1760–1774 (2012).

59. T. Smith, A. Heger, I. Sudbery, UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* . **27**, 491–499 (2017).

60. G. Gambardella, D. di Bernardo, A Tool for Visualization and Analysis of Single-Cell RNA-Seq Data Based on Text Mining. *Front. Genet.* **10** (2019), doi:10.3389/fgene.2019.00734.

61. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. **26** (2010), doi:10.1093/bioinformatics/btp616.

62. J. H. Levine *et al.*, Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*. **162**, 184–197 (2015).

63. R. L. Grossman *et al.*, Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).

64. A. Colaprico *et al.*, TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71–e71 (2015).

65. J. Barretina *et al.*, The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. **483**, 603–607 (2012).

Methods

Cell culture: The 33 cell lines used in this study were obtained from Nerviano Medical Sciences and cultured in ATCC recommended complete media at 37°C and 5% CO2.

DROP-seq platform set-up: Single cell transcriptomic of the 33 cell-lines was performed with DROP-seq platform (*17*). The microfluidics device useful for the generation of droplet was fabricated using a bio-compatible, silicon-based polymer, polydimethylsiloxane (PDMS) that was rendered hydrophobic with Aquapel® treatment. In each sequencing experiment, cell suspension, bead suspension and carrier oil (QX200 droplet generation oil, Bio-Rad) were first loaded in syringes and then placed in syringe pumps (Leafluid). Flow rates of syringe pumps were set at 4,000 μL/hr for both cell and barcoded bead suspensions while carrier oil syringe pump was set at 15,000 μL/hr. In each sequencing experiment, cells and barcoded beads were respectively diluted at the concentration of 200 cell/μL in PBS with BSA 0.01% (Merck) and 120 bead/μL in lysis buffer. A self-built magnetic stirrer system was used to keep in suspension barcoded beads. To count the occurrence of a single cell together with a barcoded bead several tests were performed without lyses buffer in the bead suspension. In these tests we observed about 5% of generated droplets filled with just one bead and one cell.

Single cell RNA library preparation and sequencing: In each sequencing experiment the targeted number of cells to sequence was set to 2,000. Droplets were collected in a 50 mL falcon and broke by adding 1 mL of Perfluoro-1-octanol. Captured RNA was reverse transcribed in a single reaction following the original protocol described in (*17*) and then digested with exonuclease 1 to degrade unbound primers. Next, cDNA was first amplified with a total of 12 PCR cycles and then purified using AMPure XP beads at 0.6X ratio. Finally, the quality of the resulting cDNA library was quantified on BioAnalyzer High Sensitivity DNA Chip and its concentration measured using the Qubit Fluorometer. Then the Illumina Nextera XT v2 kit was used to produce the next generation sequencing (NGS) libraries using four aliquots of 600pg of each cDNA library. Quality and concentration of NGS libraries were respectively quantified on the BioAnalyzer High Sensitivity DNA Chip and Qubit Fluorometer. Finally, either Illumina NextSeq 500/550 or NovaSeq 6000 machines were used to sequence the produced NGS libraries (Supplementary Table 01). Samples processed with NextSeq500/550 NGS library were diluted at the final concentration of 3 nM and sequenced using the 75 cycle high output flow cell while samples processed with NovaSeq 6000 machine were diluted at the final concentration of 250 pM and sequenced using the S1 100 cycles flow cell.

Read alignment and gene expression quantification: Raw data processing was performed using the Drop-seq tools package version 1.13 and following what is described in the Drop-seq Core Computational Protocol (http://mccarrolllab.org/dropseq). Briefly, raw sequence data was filtered to remove all read pairs with at least one base in their barcode or UMI with a quality score less than 10. Then read 2 was trimmed at the 5' end to remove any TSO adapter sequence, and at the 3' end to remove polyA tails. Reads were then aligned using STAR (*57*) on hg38 human genome (primary assembly, version 28) downloaded from GENCODE (*58*). After reads alignment, UMI tool (*59*) was used to perform UMI deduplication and quantify the number of gene transcripts in each cell. The initial number of sequenced cells was identified using a simple (knee-like) filtering rule as implemented by CellRanger 2.2.x. After this, only high depth cells with at least 2,500 UMI, more than 1,000 captured genes and with less than 50% of reads aligned on mitochondrial gene were retained. Putative multiples among the sequenced cells of each BC cell-line were simply discarded identifying outliers in the count depth distribution by using Tukey's method based on lower and upper quartiles with k equal to 3.

BC Atlas Construction: Single cells expression profiles were normalized using GF-ICF (Gene Frequency – Inverse Cell Frequency) normalization we recently developed into the *gficf* package (*60*) for R statistical environment (https://github.com/dibbelab/gficf). GF-ICF is based on a data transformation model called term frequency-inverse document frequency (TF-IDF) that has been extensively used in the field of text mining. Briefly, given a set of N cells, let f_{ij} be the number of transcripts of the gene i in the cell j, the gene frequency

GF_{ij} of gene i in the cell j, can be defined as: $GF_{ij} = f_{ij} / \sum_{k=1}^{N} f_{kj}$ and represents its number of transcripts divided by the total number of transcripts of the cell. The Inverse Cell Frequency of gene i can be instead defined as $ICF_i = \log(N + 1/n_i + 1)$ where n_i denotes the number cells that contain gene i among the N sequenced cells. The GF-ICF score for gene i in cell j is finally defined as $GF_{ij} \times ICF_i$. GF-ICF values of each cell are then rescaled to have Euclidean norm equal to one (L2 normalization) to account for cell depth biases. GF-ICF transformation was applied on CPM (count per million) after *EdgeR* normalization (*61*) and discarding genes expressed in less than 5% of the total number of sequenced cells. Finally, each cell was summarized with its first 10 Principal Components (PCs) and projected with UMAP (*20*) into a two dimensional embedded space. The number of principal components was choose looking the elbow plot of the first 50 PCs. UMAP projection was performed by using the *uwot* package in the R statistical environment 3.6.

Cell clustering and identification of marker genes: Transcriptionally similar subpopulations of cells were found using a Phenograph like approach (*62*) as implemented in the *clustcells* function of *gficf* package. Briefly, we intially built a graph of cells by using the K-Nearest Neighbours (KNN) algorithm applied on the PC-reduced space where each cell was connected to its 50 most similar cells using the manhattan distance. Then, to build the final graph of cells edge weight between any two cells was refined with Jaccard similarity by computing the proportion of neighbours they share. Louvain algorithm with resolution parameter equal to 0.25 was finally used to find communities of cells in this graph of cells. Differentially expressed genes in each cluster were instead identified using *findClusterMarkers* function of *gficf* package that compare the expression of a gene in each cluster versus all the other by using the Wilcoxon rank-sum test.

BC bulk expression dataset: Raw BC bulk expression data and relative patient clinical information were collected from the Genomic Data Commons (GDC) portal (*63*) by using the *TCGAbiolinks* package (*64*). Then, raw counts were normalized using the *EdgeR* package (*61*) into R statistical environment 3.6. and converted in log10(CPM+1) before to be plotted and any other analysis.

Spatial sequencing data: Spatial transcriptomic data of BC patient was download from 10x Genomic website (https://www.10xgenomics.com/resources/datasets). Only tiles reported to be "in tissue" according to the related metadata of each patient slide were used.

Embed new cells into the BC atlas and prediction of the cancer type: New points were embedded into the UMAP space via *embedNewCells* function of *gficf* package. Briefly, tiles from 10x spatial transcriptomic were normalized with GF-ICF method as described above but using the ICF weight estimated on the BC atlas. Then tiles are projected in the existing PC space using gene loadings estimated on the BC atlas. After this transformation, tiles were embedded in the BC atlas via *umap_transform* function of *uwot* package. Finally the cancer type of each new embedded point was predicted using the function *classify.cells* of the package *gficf* with the k parameter equal to 7. This function performs a k-nearest neighbour classifier to classify the new embedded points using the coordinates of the UMAP space.

Single-cell drug sensitivity prediction: The basal expression profile of about 1,000 cancer cell line was obtained from RNA-sequencing data downloaded from the Cancer Cell Line Encyclopaedia (CCLE) portal (*65*). Cell line belonging to liquid tumour were discarded and only 658 cell lines belonging to solid tumours were retained and used for further analysis. The raw counts of each gene were then normalized with edgeR package (*61*) and transformed in log10(CPM+1). Lowly expressed genes and genes whose entropy was in the fifth percentile were excluded from the analysis. Expression profiles of the 658 CCLs were then crossed with drug sensitivity data from work of Rees and colleges (*13*). This dataset was originally composed by 481 small molecules, but after removing drugs for which the in vitro response was available for more than 25 CCLs only 450 small molecules were retained for further analysis. Then, for each gene and for each of the 450 drugs we computed the Pearson correlation coefficient (PCC) between the expression of the gene and the effect of the drug expressed in terms of Area Under the Curve (AUC) across the 658 cell-lines. Since the AUC reflects the in vitro response of a cell-line to different concertation of a drug in a timeframe of 72 hours, lower values of AUC are associated with sensitivity whereas higher values with resistance to the drug by the

tested cell-line. Hence, genes positively correlated with the AUC are potential markers of resistance (the more expressed the gene, the higher the concentration needed to inhibit growth), vice-versa, negatively correlated genes are markers of sensitivity. We this approach, we generated a ranked list of expression-based biomarkers of drug sensitivity and resistance for each of the 450 drugs where genes positively correlated with the AUC are at the top, and those negatively correlated at the bottom. Finally, to predict drug sensitivity at the single-cell level, we used the top 250 expressed genes of each cell as input of Gene Set Enrichment Analysis (GSEA) (*51*) against the ranked list of biomarkers for each one of 450 drugs built as described above. Hence, while a negative enrichment score implies that genes associated to drug sensitivity are highly expressed by the cell, a positive one indicates the cell express genes conferring her drug resistance. GSEA and associated p-values were estimating using the *fgsea* package in the R statistical environment version 3.6.

Validation of drug sensitivity prediction: Precision of the proposed method in predicting drug sensitivity was evaluated using an independent drug screening dataset produced from Iorio and colleagues (*12*) composed by 1,001 CCLs and their maximal inhibitory concentration (IC50) values at 265 small molecules. Hence, we used the described method on our 32 BC cell-lines to predict from single cell transcriptional data the percentage of sensitive cells at 86 common drugs between the two datasets. The "golden standard" was built by assigning to each of 32 x 86 (=2,752) cell-line/drug pair the value 1 if the cell line was sensitive to the drug and 0 otherwise. To determine if a cell-line was sensitive or not to a specific drug, we converted for each drug its IC50 distribution in Z-scores using all the 1,001 available cell-lines and then defined a cell-line sensitive to the drug if and only if its Z-score was in the 5% percentile. Finally, Positive Predicted Values (PPV) were defined as TP/(TP+FP) where TP represents the number of true positives and FP the number of false positives predicted cell-lines/drug pairs.

Prediction of cell cycle phase from scRNA-seq: The cell cycle phase of each sequenced cell was predicted using the function *CellCycleScoring* of the *Seurat* tool with default parameter and following what was suggested in the corresponding vignette (https://satijalab.org/seurat).

HER2 antibody staining procedure for flow cytometry analysis: Before staining**, c**ells were first washed with phosphate-buffered saline (PBS) 1x, detached with 0.05% trypsin-EDTA, resuspended and harvested with the appropriate medium in single-cell suspension. Then, cells were counted, washed with PBS-FBS 1% and finally incubated for 15 min at 4° in the dark at the concentration of 1.0×10^6 cell/μL with staining buffer. The staining buffer was prepared diluting the mouse anti-human HER2 antibody (BD BB700) at the final concentration of 0.00114 ng/μL. Then, to remove unbound antibody, cells were washed three times with PBS-FBS 1%. Flow cytometry measurements were performed on either BD Accuri C6 or BD FACSAria III instruments. In order to define antibody positive and negative cells, the unstained samples were used to set the gate. To record data, at least 2.0×10^4 events were collected for each sample. Data analysis were performed using the BD Accuri C6 software.

HER2 expression dynamics experiment: Sorting of MDA-MB-361 HER2-positive and HER2-negative cells was performed following the antibody staining procedure described above with the only exception that before sorting, each sample was resuspended in sorting buffer (PBS 1x, FBS 1%, trypsin 0.1%, EDTA 2mM). Then, 4.0×10^5 cells were collected for each cell subpopulation (*i.e.* HER2-positive and HER2-negative), plated in their appropriate medium and incubated at 37°. After 18 days, the percentage of cells expressing HER2 protein was checked by performing the antibody staining procedure described above.

Drug sensitivity assay: Cells were seeded in 96-well microplates (PerkinElmer); the seeding cell confluency was specifically optimized for each cancer cell line to have cells in growth phase at the end of the assay. After overnight incubation at 37°, cells were treated with DMSO (Merck) for the negative control and with five concentrations of selected drugs in triplicate. Cells were then incubated at 37° for 72hr. Cell viability was assessed by measuring either luminescence with GloMax® Discover instrument from Promega or by nuclei count using the Operetta instrument from PerkinElmer. For luminescence measurement, cells were treated with Promega CellTiter-Glo® Luminescent Cell Viability Assay according to the manufacturer protocol. For nuclei count, cells were washed with PBS 1x, fixed with paraformaldehyde (PFA) 4% for 10 min at room temperature, washed with PBS 1x, incubated at room temperature in the dark with HOECHST

33342 (Thermo Fisher Scientific) diluted 1:1000 in PBS 1x for 10 min and finally washed with PBS 1x. Nuclei count was performed using Columbus image analysis software (PerkinElmer). All drug used in this study were purchased from Selleckchem.

Data availability:

Code availability:

Acknowledgments: This work was supported by the STAR (Sostegno Territoriale alle Attività di Ricerca) grant of University of Naples Federico II and the AIRC (Associazione Italiana Ricerca sul Cancro) GRANT MFAG 23162 to GG and by the AIRC (Associazione Italiana Ricerca sul Cancro) Grant IG 2016-18479 to DB.

Author Contribution: GG performed all computational analysis, conceived the method for single-cell drug sensitivity prediction and contributed to the writing of the manuscript. GV implemented the dropseq platform, performed single-cell RNA sequencing and drug response validations. BT performed FACS analyses, helped with cell culture and RNA-seq library preparation. AI and RB provided the breast cancer cell-lines and contributed to the writing of the manuscript. DdB supervised the work, contributed to the writing of the manuscript and conceived the original idea.

Conflicts of interest

The Authors declare no conflict of interests.

Author Approvals: All authors have seen and approved the manuscript. The manuscript has not been accepted or published elsewhere.