Single-cell transcriptomics reveals multi-step adaptations to endocrine therapy

Sung Pil Hong1,2,7, Thalia E. Chan3, Ylenia Lombardo1,6, Giacomo Corleone1, Nicole Rotmensz4, Sara Bravaccini5, Andrea Rocca5, Giancarlo Pruneri4, Kirsten R. McEwen3, R. Charles Coombes1, Iros Barozzi1,7 & Luca Magnani1

Resistant tumours are thought to arise from the action of Darwinian selection on genetically heterogeneous cancer cell populations. However, simple clonal selection is inadequate to describe the late relapses often characterising luminal breast cancers treated with endocrine therapy (ET), suggesting a more complex interplay between genetic and non-genetic factors. Here, we dissect the contributions of clonal genetic diversity and transcriptional plasticity during the early and late phases of ET at single-cell resolution. Using single-cell RNA-sequencing and imaging we disentangle the transcriptional variability of plastic cells and define a rare subpopulation of pre-adapted (PA) cells which undergoes further transcriptomic reprogramming and copy number changes to acquire full resistance. We find evidence for sub-clonal expression of a PA signature in primary tumours and for dominant expression in clustered circulating tumour cells. We propose a multi-step model for ET resistance development and advocate the use of stage-specific biomarkers.
Absence of features of resistance in treatment-naive cells

These results highlight the multi-faceted effects of ET at single-cell mesenchymal traits, which is found dominant in clusters of circumscribed transcriptional clones with features of resistance. To investigate this, we generated scRNA-seq high-quality profiles for >1200 MCF7 and >1900 LTED cells (Supplementary Table 2).

Dimensionality reduction (Similarity Weighted Nonnegative Embedding, or SWNE) showed MCF7 and LTED as completely separated populations, with no single MCF7 clustering with LTED cells (Fig. 1b). Studies in melanoma and TNBC suggest that drug-resistant cells can rapidly emerge. This implies that in drug-naïve tumours, at least a few cells have a transcriptional profile similar to that of fully resistant cells. However, our data suggest this is not the case in luminal breast cancer cell lines, which is concordant with the long latency taken by resistance to occur in most patients treated with ET. To completely exclude any contribution of a pre-existent genetic clone, we inferred single-cell, copy number alterations (CNAs) from scRNA-seq data (see Methods). Clustering of single MCF7 and LTED cells based on the inferred patterns of CNAs identified two clades, one including all the MCF7 and one all the LTED cells (Fig. 1c). In line with CYP19A1 significantly contributing to AI resistance in vivo and in vitro, an amplification involving the region was found in LTED cells, but not in MCF7 (Fig. 1c). This was confirmed by shallow whole-genome sequencing (Supplementary Fig. 1a). Clustering of single-cell profiles identified five distinct groups (two for the MCF7 and three for the LTED), mainly driven by differences in cell cycle (Fig. 1d). Even after running the dimensionality reduction step separately on cells assigned to the same cell-cycle phase, MCF7 and LTED cells were unambiguously separable (Supplementary Fig. 1b). Importantly, scRNA-seq confirmed that previously reported pathways, such as cholesterol biosynthesis, are profoundly reprogrammed by ET (Fig. 1d; Supplementary Fig. 1c, d). Taken together, these data support that AI resistance is not driven by a pre-resistant clone (whether genetic or in a particular transcriptional state), suggesting a multi-step adaptation process in which the necessary hits occur with a different timing during ET. Nevertheless, we could not exclude the presence of a rare, transcriptionally defined clone at a very low frequency. This led us to leverage previously acquired knowledge on cancer cell plasticity to further dissect the phenotypic heterogeneity of cells in the drug-naïve condition.

Phenotypic heterogeneity of luminal breast cancer cells. Previous studies identified CD44 as a marker of plastic cells in various solid tumours. It has been suggested that CD44-positive cells possess increased tumorigenic ability and resilience to pharmacological treatments. To investigate the potential role of CD44 as a surface marker to guide the dissection of the phenotypic heterogeneity of luminal breast cancer cells, we identified those genes showing high transcriptional variability across single MCF7 cells (n = 778) and intersected them with annotation from the Cell Surface Protein Atlas. CD44 was indeed found among those genes showing high transcriptional variability across single MCF7 cells (n = 778) and intersected them with annotation from the Cell Surface Protein Atlas. CD44 was indeed found among those genes showing high transcriptional variability across single MCF7 cells (n = 778) and intersected them with annotation from the Cell Surface Protein Atlas.

Results

Absence of features of resistance in treatment-naive cells. In this study, we use a combination of live cell imaging, single-cell RNA-sequencing (scRNA-seq) and machine learning to dissect the phenotypic heterogeneity and plasticity of ERα-positive BCa, and leverage this information to identify a subpopulation of rare, pre-adapted cells both in vitro and in vivo. These cells (termed PA, from Pre-Adapted) display a unique transcriptional signature with features of dormancy and mixed epithelial and mesenchymal traits, which is found dominant in clusters of circulating tumour cells. PA cells show a significant survival advantage under short-term ET, but require further transcriptional reprogramming and genetic alterations to acquire full resistance and re-establish a proliferative phenotype in vitro. These results highlight the multi-faceted effects of ET at single-cell level, and suggest a multi-step mechanism of drug resistance that involve both non-genetic and genetic contributions.

Phenotypic heterogeneity of luminal breast cancer cells. Pre-
(Fig. 2b; Supplementary Fig. 3d; twofold, \( p \)-value = 0.0029; Wilcoxon signed-rank test), suggesting higher chances of survival to ET for cells expressing CD44 in vivo. We next sought to investigate if CD44\(^{\text{high}}\) cells can be also found at other active sites in breast cancer patients. Interestingly, we found substantial CD44\(^{\text{high}}\) cells in pleural effusions from all four patients examined (Supplementary Fig. 3e). In line with this, the fraction of CD44\(^{\text{high}}\) cells was significantly increased in LTED (upper panels in Supplementary Fig. 3f, g). Extensive functional characterisation of these cells demonstrated that MCF7-CD44\(^{\text{high}}\) cells were more invasive, more clonogenic and could form first- and second generation of mammosphere at higher efficiency than CD44\(^{\text{low}}\) cells (Supplementary Fig. 3j–l). In agreement with previous studies\(^{35}\), CD44\(^{\text{high}}\) cells also showed cellular plasticity as they could recapitulate the entire population, while CD44\(^{\text{low}}\) were capable of generating only CD44\(^{\text{low}}\) cells (Supplementary Fig. 3f).

To further investigate the plasticity of CD44\(^{\text{high}}\) cells in vitro at the single-cell level, we generated MCF7 and LTED cell lines with a GFP reporter expressed under the promoter of the CD44 gene (Supplementary Fig. 4). Reconstitution experiments from sorted cells showed that CD44\(^{\text{GFP-high}}\) cells could recapitulate all the functional aspects of endogenous CD44\(^{\text{high}}\) cells, including cellular plasticity (Fig. 2c). Interestingly, both CD44\(^{\text{GFP-high}}\) and CD44\(^{\text{GFP-low}}\) showed features of plasticity in fully resistant cells.
(Fig. 2d; Supplementary Fig. 3g). When MCF7 were challenged with short-term ET, only CD44\textsuperscript{GFP-high} cells appeared to adapt to it, while CD44\textsuperscript{GFP-low} cells were rapidly cleared out between days 4 and 7 (Fig. 2e). Single-cell plating experiments confirmed that only CD44\textsuperscript{high} cells could drive the formation of early colonies under E2 deprivation, but the colonies were significantly smaller compared with E2-supplemented conditions (Fig. 2f). These observations indicate combined cytostatic and cytotoxic effects of ET and that those cells that could adapt to the therapy originate within the CD44\textsuperscript{high} compartment. Extrapolation of cell-cycle dynamics of CD44\textsuperscript{GFP-high} and CD44\textsuperscript{GFP-low} cells from time-lapse imaging data revealed comparable cell-cycle length in E2-supplemented condition (Fig. 2g; +E2). Nevertheless, CD44\textsuperscript{GFP-high} cells had a significantly lower proportion of cells engaged in productive cell division entry, suggesting the existence of a low-proliferative subpopulation within the CD44\textsuperscript{high} compartment even under permissive environments. Under E2 deprivation, the CD44\textsuperscript{GFP-low} completely failed to undergo cell-cycle entry, while 12% of CD44\textsuperscript{GFP-high} managed to do one or more cell cycles, with a much longer latency (Fig. 2g; −E2).

Taken together, these results further support the idea that at least some of the cells in the CD44\textsuperscript{GFP-high} (but not CD44\textsuperscript{GFP-low}) compartment have an increased ability to survive the acute phase of ET, and this correlates with their features of plasticity. This led us to hypothesise that non-genetic, transcriptional variability would reflect pre-existent, rare subpopulations in treatment-naive cells with higher chances to survive and give rise to fully resistant cells.

**Transcriptional heterogeneity of plastic cells.** To investigate the transcriptional variability of CD44\textsuperscript{high} cells, we carried out sorting driven, scRNA-seq of CD44-GFP luminal breast cancer cells. About 10,000 single cells in E2-supplemented condition were profiled (CD44\textsuperscript{GFP-high} and CD44\textsuperscript{GFP-low} in equal proportions; Fig. 3a; in the remainder of the text, these two sorted subpopulations will be referred to as CD44\textsuperscript{high} and CD44\textsuperscript{low}). Dimensionality reduction (Fig. 3a) highlighted a surprising similarity between the profiles of CD44\textsuperscript{high} and CD44\textsuperscript{low}, except for a small percentage (~4%) of CD44\textsuperscript{high} cells significantly departing from the main cluster. In line with this, differential expression analysis of the two subpopulations resulted in tenfold less differentially expressed genes (DEGs) than those observed by comparing them to LTED (Fig. 3b; Supplementary Data 1). Nevertheless, CD44\textsuperscript{high} showed an overall, significantly higher transcriptomic variability (p-value < 2.2e-16; Wilcoxon rank-sum test) than CD44\textsuperscript{low} (Fig. 3c).

We next sought to systematically address whether the observed variability was the result of either an increased transcriptional noise specific to CD44\textsuperscript{high} cells (compatible with a bet-hedging mechanism) or instead the reflection of a regulated network (leading to coordinated expression of multiple genes in the same cell). We applied PIDC\textsuperscript{37}, an algorithm using partial information decomposition (PID), to identify regulatory relationships between genes, and reconstructed the gene regulatory networks (GRNs) from the scRNA-seq profiles of CD44\textsuperscript{high} and CD44\textsuperscript{low} cells, separately (Supplementary Data 2). The two networks were merged and analysed to identify major communities (Fig. 3d; the three largest communities were consistently identified on the separate CD44\textsuperscript{high} and CD44\textsuperscript{low} networks, with >95% overlap with the corresponding community from the merged network; Supplementary Table 3). The largest of the three identified communities (#1 in Fig. 3d–f) showed the lowest similarity between the CD44\textsuperscript{high} and CD44\textsuperscript{low} GRNs, with the majority of edges supported only by the CD44\textsuperscript{high} GRN (Jaccard Index = 0.23, still higher than expected by chance, expected value = 0.0154,
Single-cell transcriptomics identifies pre-adapted cells. To investigate the role of transcriptomic variability of plastic cells during acute-ET, we performed scRNA-seq experiments upon oestrogen deprivation (Supplementary Table 2). Continuous single-cell imaging suggested that cells within the CD44^low subpopulation started being differentially affected by acute-ET after 48 h of treatment (Fig. 2e). We thus profiled gene expression data of about 10,000 single cells at 48 h of E2 starvation (Fig. 4a). Applying a stringent threshold on the first SWNE component, we could define a rare, pre-adapted (PA) subpopulation among plastic cells (CD44^high) expressing a signature of acute-ET even in permissive E2-supplemented condition. The identification of PA cells was confirmed using an orthogonal approach aimed at identifying outliers and based on Random Forests classification (Fig. 4b, Supplementary Data 3 and Methods). We then excluded the PA cells identified using SWNE (Fig. 4a), and trained another Random Forest classifier that was tested on the PA cells (Supplementary Fig. 5a). Overall, 72.8% of PA cells were mis-classified as starved cells, compared with an expected 2.2% (out-of-bag classification error), further corroborating the observation that PA cells are strongly biased towards features of starved cells. Of note, PA cells are genetically indistinguishable from the other CD44^high cells, and have not yet acquired any of the genetic rearrangements of the fully resistant, LTED cells (Fig. 4c). Considering both approaches and either a lenient or a stringent threshold, PA cells are estimated to constitute 0.76–4% of the CD44^high cells, which correspond to 0.03–0.14% of the total MCF7 population. Overall, these data suggest that PA cells might represent the first step in the process of adaptation to acute-ET.

We then sought to validate if the PA transcriptional state would confer a survival advantage compared with other plastic cells exposed to acute-ET. First of all, we identified the Claudin-1 gene (CLDN1) as a suitable surface marker to enrich for PA cells by FACS in combination with CD44 (Supplementary Data 3 and Supplementary Fig. 5b). We then generated MCF7 cells stably labelled with either a nuclear GFP or mKate2 and leveraged this tool to follow two subpopulations over time after mixing them. The same amount of sorted PA cells (CD44^high and CLDN1^high) was mixed with other plastic cells (CD44^high and CLDN1^low, Fig. 4d). CD44^high CLDN1^high PA cells showed increased survival to acute-ET compared with CD44^high CLDN1^low, with this effect increasing over time. As a control, no difference was observed between CLDN1^high and CLDN1^low from the CD44^low compartment. These data strongly support the hypothesis that PA cells have a distinctive survival advantage under acute-ET (Fig. 4e).

We then further characterised these cells functionally, focusing on the set of differentially expressed genes between the PA cells and the rest of the CD44^high cells in +E2 condition (cells identified through the SWNE-based approach; 312 upregulated and 1242 downregulated; Fig. 4a, b; Supplementary Data 3). PA cells displayed features of mixed epithelial and mesenchymal traits, along with upregulation of p53 pathway, cell polarity (apical junction components) and hypoxia (Fig. 5a, upper panel). PA cells also showed reduced ERα activity and downregulation of the cell cycle machinery, while still expressing ESR1 (Fig. 5a, lower panel and Supplementary Fig. 5c). Interestingly, both plastic and non-plastic cells lied on a continuum showing a...
negative correlation between the expression of the genes of the cell cycle and of those in the signature of PA cells (Fig. 5b; Spearman’s rank correlation coefficient = −0.519; p-value < 2.2e−16), with PA cells found at the edge of this spectrum. We finally sought to quantify the overlap between the PA cells signature (upregulated genes) with the CD44* enriched GRN we previously identified (Fig. 3d–f). Indeed, when we further dissected the GRN (community #1) into its two main components, we found extensive overlap between one of these components and the PA signature (Fig. 5c; p-value = 2.7e−21; hypergeometric test). This further supports the idea that the genes in this signature are part of a co-regulated network.

Overall, these data support the hypothesis that plastic cells are phenotypically heterogeneous (with no evidence supporting genetic clones), and that among them rare cells in the PA state have a survival advantage during acute-ET.

**PA features persist in acute-ET, but not in full resistance.** While these analyses support a pivotal role for the PA phenotype in conferring a survival advantage during acute-ET, PA cells are still genetically indistinguishable from the rest of the cells. This suggests these cells do not represent the final step of drug resistance. Nevertheless, we aimed at determining whether longer exposure to acute-ET correlates with the persistence of the PA signature, and/or this also coincides with other reprogramming events. In order to capture the different dynamics of survival of CD44* and CD44* cells (Figs. 2e, 6a), we generated scRNA-seq profiles at 4 and 7 days of E2 deprivation (Supplementary Table 2), a period in which the relative number of CD44* cells does not change while CD44* cells undergo rapid extinction. Dimensionality reduction of ≥28 k cells showed increased prevalence of the PA signature with time of starvation (Fig. 6a).

Formal quantification using AUCell confirmed this trend
The same analysis using a LTED-specific signature (Methods and Fig. 1) failed to identify any cell expressing it during acute-ET (Fig. 6b, right panel). In line with this, the critical transcriptional pathways driving full resistance (i.e., cholesterol biosynthesis and re-activation of ERα signalling) were completely abrogated in PA and cells exposed to acute-ET (Fig. 6c, d). On the other hand, some of the pathways associated to PA phenotype (partial-EMT, cell polarity, hypoxia) were found to consistently increase during treatment.

Unexpectedly, while imaging showed that after 7 days >75% of the CD44low died and were destined to extinction (Fig. 2f), the profiled CD44high and CD44low cells converged on the same transcriptional changes. We reasoned that since scRNA-seq experiments capture viable cells exclusively, we profiled only those cells that were still alive at day 7. Thus, we hypothesised that the PA-like transcriptional programme is an intermediate bottleneck during acute-ET. In line with this, we discovered that CD44low cells can occasionally upregulate a signature overlapping that of PA cells, but this happens with lower efficiency (Fig. 6b), and it is not sufficient to give them the survival advantage shown by CD44high. To validate these observations at the protein level, we performed a multi-marker tracing profile, exploiting some
marker genes (namely, GPRC5A, MFGE8, FSCN1 and RAB11-FIP1) showing a trend of upregulation with starvation time. This trend was confirmed at the protein level, with values consistently higher in CD44high compared with CD44low cells (Fig. 6e; Supplementary Figs. 6, 7). Nevertheless, this did not prevent cells in the CD44low compartment to die at an almost linear rate (Fig. 6e).

Taken together, these observations confirmed that the PA transcriptome is strongly selected by acute-ET. Nevertheless, the observed rapid expansion (Fig. 6b) seems incompatible with the strict selection of a pre-existing population30. The observation that also CD44low cells can adopt a similar transcriptional profile in response to ET (despite being unable to survive) suggests that the PA programme is required, but not sufficient to explain the survival of plastic cells to acute-ET (see Discussion).

The PA signature is enriched in clusters of CTCs. T47D cells are another widely used model of hormone-dependent breast ductal carcinoma that differ from MCF7 by their TP53 status (mutated in T47D). We first confirmed that CD44 is a bona fide marker of plasticity also in these cells (Supplementary Fig. 3h, i). We then derived treatment-naive T47D cells stably expressing a GFP reporter under the promoter of the CD44 gene. With this tool, we generated high-quality scRNA-seq profiles from sorted T47D-CD44high cells, either in the presence or absence of E2. We also profiled unsorted population of T47D and LTED cells (capturing ~3000–4000 cells each; Supplementary Table 2). In line with what observed with MCF7 (Fig. 1b), dimensionality reduction (SWNE) indicated that no single treatment-naive cell clustered with LTED cells (Fig. 7a; Supplementary Fig. 8a; Supplementary Data 4). We then looked specifically to CD44high cells and, similarly to MCF7 (Fig. 4a), we were able to identify a small fraction of treatment-naive cells overlapping with the E2-deprived cells (Fig. 7b). Up- and downregulated genes in T47D-PA cells showed extensive, highly significant overlap with those single out in MCF7-PA (Fig. 7c; \(p = 2.2e-16\), hypergeometric test), with overlapping genes showing significantly higher effect size compared with the rest (Fig. 7c; \(p < 1e-5\); Wilcoxon rank-sum test). Importantly, CNAs estimated from scRNA-seq data sets support the idea that also T47D-PA cells are transcriptional clones. Although these cells do not show genetic lesions of LTED cells (Supplementary Fig. 8b), the estimated profiles are qualitatively more heterogeneous than those of the other plastic cells (Supplementary Fig. 8c). These results further strengthen a role for the identified PA signature in the survival to A1.

We then looked for evidence of expression and co-regulation of genes upregulated in PA cells, with the same genes also present in T47D-CD44high cells, in 825 primary luminal breast

---

**Fig. 6** Features of pre-adaptation persist in acute-ET, but not in full resistance. a Sampling design along with dimensionality reduction of single-cell transcriptional profiles of E2 supplemented (day 0) or deprived (days 2, 4 and 7). b AUCell quantification of the fraction of single-cells showing transcriptome compatible with either the pre-adapted (left) or the LTED (right) signatures. c Selected gene set enrichment across all conditions profiled in this study. d Score distributions for the indicated gene sets, across cells. e Multi-marker tracing profiles for selected genes (box plots) in CD44high and CD44low cells upon E2 deprivation. Survival (as relative number of residual cells) is also shown (bar plots). Box plots show median, interquartile values, range and outliers (individual points).
Tumours classified as luminal A showed significantly higher expression of the signature compared with luminal B (p-value < 2.2e-16; Wilcoxon rank-sum test) and TNBC/HER2+ lesions (p-value = 1.9e-8) (Supplementary Fig. 8d). Of note, luminal A exhibits the longest latencies in relapse development amongst all BCa41–43. Considering > 600 luminal A samples, we then checked the distribution of pairwise correlations between the expression pattern of the genes in the signature, as a proxy for co-regulation. Compared with a size-matched set of randomly picked genes, those in the PA signature showed significantly higher coefficients (Supplementary Fig. 8e; p-value < 2.2e-16; Wilcoxon rank-sum test), with hundreds of pairs with values over 0.5 (Spearman’s rank correlation coefficient). These results further corroborate our previous observations that these genes tend to be controlled by the same GRNs, and showed a trend of higher expression in luminal tumours with longer latency of recurrence (A vs B; Supplementary Fig. 8d).

Given that some of the key pathways active in PA cells hinted to mixed epithelial and mesenchymal features, as well as cell polarity and migration, we asked if the PA phenotype could play a role in metastatic progression. Previous data strongly suggest that epithelial-like clusters of circulating tumour cells (CTCs) are responsible for 85–92% of metastatic dissemination44, with individual CTC showing more mesenchymal features playing a
more limited role\textsuperscript{45}. Interestingly, the PA signature was found significantly enriched in CTCs\textsuperscript{45} (Fig. 7d; \( q\)-value = 0.017, permutation test) and at even higher levels in clusters of CTCs\textsuperscript{44} (Fig. 7e; \( q\)-value = 0.066, permutation test). These results provide a further link between drug-induced adaptation and metastatic invasion\textsuperscript{27,46}.

**Discussion**

In this study, we leveraged two in vitro models to investigate the contribution of genetic and transcriptional heterogeneity to the development of resistance to ET in luminal breast cancer. As opposed to previous observations in melanoma, TNBC, lung and colorectal cancers, in which targeted therapy lead to the rapid emergence of fully resistant cells\textsuperscript{10–12,18,26}, we could not find any genetic or phenotypic clone showing features of resistance in treatment-naive cells (Figs. 1, 7). The same observation held true even after thoroughly dissecting the heterogeneity of the cells showing features of plasticity (Figs. 2, 3 and 7). On the other hand, we could identify and characterise a small subpopulation (\(-0.1\%\) of the treatment-naive cells) showing a PA phenotype (Figs. 4, 7). These cells showed a twofold increased survival to acute-ET compared with other plastic cells (while non-plastic cells undergo complete extinction under selective pressure; Fig. 4e), along with mixed epithelial and mesenchymal features, and quiescence. Interestingly, while any cell (also those with no feature of plasticity) can adopt a transcriptional programme overlapping that of the PA cells, only plastic cells can withstand acute-ET (Fig. 6a, e), with PA cells showing a more pronounced survival advantage (Fig. 4e). Finally, we found an enrichment of the PA signature in clusters of CTCs, linking a quiescent subpopulation from the primary tumour to both features of survival to therapy and of CTCs. Interestingly, it has been reported that early-stage metastatic cells possess partial features of survival, dormancy and EMT, which all overlap with our PA signature\textsuperscript{47}. A signature of partial-EMT has also been recently shown to be expressed in the cells at the leading edge of primary head and neck cancers\textsuperscript{48}. It is tempting to speculate that PA cells might not only display a survival advantage during the early phases of the therapy but might also be the pioneers of micro-metastatic spread.

Surprisingly, we found that also cells with no features of plasticity were able to adopt the PA signature, even though with a much lower efficiency, which cannot prevent the extinction of the compartment after two weeks of E2 deprivation (Fig. 6). On top of this, 70\% of the plastic cells adopted a PA signature within 48 h of acute-ET (Fig. 6b). This fast transition to a diverse transcriptional state is hardly explained by conventional Darwinian selection of a pre-resistant (or persister) cell\textsuperscript{30}. For reasons that remain to be investigated, plastic cells have a much higher probability than non-plastic ones to transition into a PA state, and this probability is dramatically increased by E2 deprivation. We reason that upon stress, plastic PA cells are better positioned than cells requiring transcriptional reprogramming, hence the observed difference in survival within the plastic compartment (Fig. 6e). We estimated PA cells to constitute ~0.1\% of the treatment-naive cells. In order to obtain ~100 PA cells would have required profiling at least 70,000 MCF7 cells by scRNA-seq. Even in the best-case scenario, this single experiment requires capturing more cells than those profiled across this entire study (Supplementary Table 2). This suggests that functional approaches leading to dissection of the phenotypic heterogeneity, and thus to enrichment strategies (Fig. 2) are required for the feasibility of this kind of studies.

The data presented here suggest PA cells as an obligated step towards the acquisition of resistance while still requiring substantial reprogramming to recapitulate features of fully resistant cells (Fig. 8). We propose that the delayed relapse common to ET-treated patients might be mediated by similar processes, in which PA-like cells are selected for and stalled by ET for up to >10 years. This model would reconcile why ET are sometime effective for downstaging neo-adjuvant patients, but fail to clear micro-metastatic disease. Nevertheless, single-cell lineage-tracing approaches coupling unambiguous identification of clones to transcriptome mapping are needed to get a definitive proof that it is the progeny of PA cells that will eventually acquire full resistance. Besides, how this bottleneck affects the progression of the tumour requires further investigation. Future studies on the necessary steps and their timing of occurrence during treatment must be carried out in order to expose potential vulnerabilities of these quiescent cells.

**Methods**

**Cell lines.** MCF7 and long-term oestrogen-deprived cells (LTED) were kindly provided by Philippa Darbre and T47D and LTED were kindly provided by Matthew Ellis\textsuperscript{49}. MCF7 and T47D cells were maintained in the Dulbecco’s modified Eagle’s medium (DMEM) containing 10% foetal calf serum (FCS) Long-term oestrogen-deprived cells (LTED) were derived from MCF7 or T47D after 1 year oestrogen deprivation and were maintained in phenol-red free DMEM containing 10% charcoal stripped foetal calf serum (FCS)\textsuperscript{50}. Both media were supplemented with 2 mM L-glutamine, 100 units/mL penicillin and streptomycin. \( 10^{-8} \) M oestradiol (E2758 Sigma) was added routinely to MCF7. Primary-metastatic breast cancer cells were derived from pleural effusions of patients with metastatic breast cancer.
cancers. The pleural effusion (PE) cells were maintained in the DMEM containing 10% foetal calf serum (FCS) and 2 mM L-glutamine, 100 units/mL penicillin and streptomycin. Written informed consent for the procedure was obtained from all patients. The study was reviewed and approved by Imperial College Healthcare NHS Trust Tissue Bank (R14059). Cells were tested for mycoplasma contamination before the experiments and showed negative results.

**Plasmids.** pLVX-IRES-mCherry-puro lentiviral vector (Cambridge Bioscience, Cambridge, UK) was used to infect MCF7 and LTED cells. MCF7- and T47D-CD44 reporter GFP cells were established with CD44CR1-IRES-GFP-puro lentiviral vector (Tebu-Biochem). Stable and polyclonal cell populations were established after puromycin selection (0.5 μg/mL). NucLight Green lentivirus (Incucyte, 4626) and NucLight Red Lentivirus (Incucyte, 4627) were used to infect MCF7. Stable and polyclonal cell populations were established after Zeocin selection (300 μg/mL). The CD44 reporter GFP cells. Stable and polyclonal cell populations were established after sorting.

**Antibodies.** Anti-ErbA antibody (Vector Laboratories, VP-E613) 1:100 for immunofluorescence (IF) and anti-ErbA (Santa Cruz, HC-20) 1:1000 for western blot (WB), anti-CD44 antibody (Santa Cruz, sc-7297) 1:200 for IF and 1:100 for immunohistochemistry (IHC), anti-pan Cytokeratin antibody (Abcam, ab17154) 1:200 for IF, anti-FGFRII antibody (Abcam, ab44971) 1:1000 for WB, anti-FSCN1 (Sigma, HPA005723) 1:1000 for IF, anti-MEGF8 (Sigma, HPA002807) 1:100 for IF, anti-RAB11FIP1 (Sigma, HPA023904) 1:100 for IF, anti-GPRC5A (Sigma, HPA097782) 1:100 for IF and anti-caspase3 (Merck, MAB05733) 1:100 for IF

**FACS analysis.** Cells were cultured to 70–80% confluence and detached from the cell culture flasks using EDTA. Cell pellets were obtained and washed with cold phosphate-buffered saline (PBS) containing 1% FCS and 5 mM EDTA. All further steps were performed on ice and all centrifugation steps at 4 °C. Fluorochrome-conjugated monoclonal antibodies against human CD44 (FITC, BD Pharmingen; BV241, BD Pharmingen), Claudin-1 (APC, R&D systems), and their isotype controls were added to the cell suspension at concentrations recommended by the manufacturer (BD Biosciences) and incubated at 4 °C in the dark for 30 min. The labelled cells and CD44 reporter GFP cells were washed in PBS and then analysed on a FACS Aria (BD Biosciences). Gating was set to relevant isotype controls (FITC, BV241, Claudioin-1, APC, R&D systems). Cells to be assayed were suspended in 1 mL of 0.3% agar in medium. In all, 1 × 104 cells to 10 days for a cell-cycle analysis. Excitation (Ex) and emission (Em) filters sets (Chroma Technology Corporation) were as follows: CD44-CFP, 427-10 nm (Ex), 483-32 nm (Em). Micromanager 1.3 was used for acquisition of time-lapse images. All data analysis was done with scripts written in Matlab (Mathworks) or using Cell Profiler (Broad Institute) and Image (National Institutes of Health). Symmetric/ asymmetric/conversion analyses were performed on a total of 200 cells. Each cell was treated for the first three cell divisions (one cell to two cells, two cells to four cells, four cells to eight cells). Symmetric division was scored if the daughter cell matched the mother. Asymmetric was scored if the daughter cell did not match the mother. Division was scored if cell changed CD44 status without cell division (at least 4 h pre- or post division). Cell-cycle speed was established by calculating the time intervening between two consecutive metaphase plates.

**Statistical analyses.** Unless specified otherwise, all the analyses and plots were performed in the statistical computing environment R v3 (www.r-project.org).

**Single cell preparation.** Single cells were prepared from a full population of MCF7 and MCF7-LTED or T47D and T47D-LTED. At different time points of E2 deprivation, single cells were prepared from sorted MCF7 or T47D-CD44-GFP reporter cells by the level of GFP expression. After centrifugation, single cells were washed with PBS and were resuspended with a buffer (CaCl₂ 1 mM and MgCl₂ 1 mM and 0.04% BSA) at 1000 cells/μL.

**Single-cell RNA sequencing.** Viability was confirmed to be >90% in all samples using acridine orange/propidium iodide dye with LUNA-FL Dual Fluorescence Cell Counter (Logos Biosystems, L20001). Single-cell suspensions were loaded on a Chromium Single Cell 3’ Chip (10x Genomics), and were run in the Chromium Controller to generate single-cell gel bead-in-emulsions using the 10x genomics 3′ barcoding. After 2 days of lysis, the stained cells (1 × 10⁶) were stained using the 10x in suspension protocol. Single-cell RNA-seq libraries were prepared according to the manufacturer’s protocol, and the library quality was confirmed by a Bioanalyzer High-Sensitivity DNA Kit
Single-cell RNA-seq raw data analysis. cellRanger (v2.1.1) was run on the raw data using GRCh38 annotation (v1.2.0). Output from cellRanger was loaded into R using the function load_cellranger_matrix. parseFloat from package cellranger (v1.1.0; genome = "GRCh38"). Data sets were merged according to gene names. All cells sampled were retained, except for flow-sorted CD44high and CD44low either in +E2 media or deprived (for which the 5000 cells in terms of UMIs per cell were considered. In order to robustly detect transcriptional states, a recent paper suggested to consider a coverage of at least 1500 detected genes per cell. A filter on cells showing at least 1500 detected genes per cell, and at least 5000 UMIs per cell was then applied. After that, reads mapping on mitochondrial genes were excluded. Before normalisation, a series of filtering steps were performed. To do that, data were imported in Seurat (v2.3.4) and scaled (NormalizedData function using normalisation.method = "LogNormalize", scale.factor = 10,000, followed by the ScaleData function). A filtering step was then performed based on the cumulative level of expression (the sum of the Seurat-scaled values) of three housekeeping genes (GAPDH, RPI.26 and RPI.36). Manual inspection of these values versus the number of UMIs per cell (or the number of genes with non-zero expression per cell) revealed no correlation between the two. Nevertheless, a number of cells showed very low expression for these genes. Cells showing housekeeping gene expression in the bottom 1% were then excluded from further analysis. Finally, a filter in less than 20% cells normalisation was performed using the R package Scan (v1.6.9). Raw counts were imported into a SCE object using the newSCESet function; size factors were calculated using computeSizeFactors (sizes = seq(20, 250, 10)), on data pre-clustered through quickCluster.

Estimation of copy number alterations from scRNA-seq data. CNAs were estimated directly from the scRNA-seq data, using an approach similar to the one used by Patel et al.53. Only genes expressed in >25 cells were considered. A reference gene expression profile was generated based on published scRNA-seq profiles of hormone-responsive luminal cells (termed L2)54, using only the data sets obtained using a droplet-based approach. After normalising each single-cell profile based on the total number of detected transcripts to a fixed constant (10,000), a pseudo-bulk profile for the L2-cells was derived using the mean expression value of each gene across all cells.

Before running the actual CNAs quantification, all the raw scRNA-seq data sets generated in this study (after filtering, pre-normalisation) and the pseudo-bulk profile generated as described above were linearly normalised to a constant (10,000) and log-scaled (pseudo-count set to 1).

First of all, chromosome coordinates of all genes were retrieved using the biomart R package (v2.3.4; host set to "jul2015.archive.ensembl.org")55. This way, genes were sorted by chromosome coordinates. A genome-wide scan was then conducted using a sliding window of 100 genes, with a step of 10. Using the rollypap function from the zoo package in R (v1.8-3), mean value of expression in each bin was calculated for each single cell, as well as for the reference profile. The resulting genome-wide profile from each single cell was then linearly regressed against the reference estimate (using the function lm). The residuals were then considered a proxy for CNAs and plotted in the form of heat maps. Single-cell CNAs profiles were hierarchically clustered (hclust, method = "ward.D2") and shown as a circular dendrogram using circize_dendrogram from R package dendextend (v1.8.0). In case of full populations, CNAs were estimated on all the cells. In case of the identified pre-adapted cells, the same number of cells was randomly sampled from the other groups of cells.

Estimation of copy number alterations from ChIP-input RNA. Reads were aligned to the hg19 human reference genome using bowtie2 (v2.3.4.3)56. Aligned reads were converted to BAM files, sorted and indexed using samtools (v1.9)57. Duplicated reads were marked and removed using Picard MarkDuplicates (v2.1.1; REMOVE_DUPLICATES = true). Only uniquely mapped reads were retained for further analyses. Copy numbers were inferred using CNVkit tools (v0.9.4)58, as described here. The read-depth.hydro function of the CNVkit was run with the default parameters of the batch command after creating a flat reference genome as suggested in the manual using the command reference.

Dimensionality reduction and clustering. Normalised data were then imported in Seurat and scaled. Variable genes were identified using the FindVariableGenes function (mean.function = ExpMean, dispersion.function = LogVMR, x.low.cutoff = 0.01, y.cutoff = 0.01, for which the 100). Principal component analysis (PCA) was run using variable genes as input, and the top 50 components were kept. Clusters were then identified using FindClusters (resolution = 0.6). Considering only those variable genes identified, as described above (Similarity Weighted Nonnegative Embedding (SWNE)57, was applied to further reduce the dimensionality. The k parameter was estimated using FindNFactors on a subsample of 1000 cells (loss = "mse", 2-50 as range of values, with a step of 2). The choice of k is determined by randomly set 20% of the gene expression matrix as missing, followed by finding the factorisation that best imputes the missing values, minimising the mean square error. Using this parameter, normalisation matrix factorisation was then run through RunNMF (alpha = 0, init = "ica", loss = "mse"), followed by EmbedSWNE (alpha.exp = 1.25, snn.exp = 1.0, n.pull = n_4, dist.use = "IC"). For this step, the shared nearest neighbour (SNN) matrix calculated by the FindClusters function of Seurat was used.

Differential expression analysis. The two-sample Likelihood Ratio test implementation of the R function of the MAST R package (v1.4.13)69 was used to identify marker genes for a given sample or cluster. Briefly, each cell was either flagged either belonging to the sample (or the cluster) or not. Those genes identified as upregulated in the cluster at q-value <= 0.05 (Benjamini-Hochberg correction)60 and showing an area under the curve (AUC) >= 0.6 were classed as markers for the sample or the cluster. The AUC is an estimate on how accurately a certain gene predicts the sample or cluster. AUCs were calculated using the ROC R package61.

Functional enrichment analyses. For functional enrichment analyses, a selected number of gene sets was employed. The 50 Hallmark gene sets from the Molecular Signature Database (MsigDB)62 were downloaded from the MsigDB website on October 19, 2017. Gene sets from Puram et al.39 (Table S7), along with a manually curated list of Era-core target genes (BYSL, GREB1, HEY2, MPPHOSPH10, MYR, NIP7, RARA, SLC9A3R1, TFF1, XBP1) were also considered. For a given subset of cells, each gene set was scored separately as the sum of the normalised expression values of all the genes in the set. The resulting distributions were then used for statistical testing and visualisation.

Single-cell gene regulatory network inference. Networks were inferred separately for CD44high and CD44low cells, with nodes representing genes and edges representing statistical dependencies between gene pairs. For each data set, genes expressed in fewer than 20% of cells were excluded; then all possible network edges were ranked using the PIDC network inference algorithm63 implemented in NetworkInference (https://github.com/TcChandlers/NetworkInference.R), with expression data for each gene discretised independently into 6 bins of equal width; finally, a network was defined keeping the 2000 highest ranking edges. The two networks were then superimposed to form an overlapping network with edges belonging (i) only to the CD44high network, (ii) only to the CD44low network, or (iii) to both networks. Communities were identified in the overlapping network (and recursively in each community) using the label propagation method implemented in LightGraphs (https://github.com/JuliaGraphs/LightGraphs.jl). Communities, were required to include at least ten nodes. Similarity of the CD44high and CD44low networks within each community was calculated using the Jaccard index: the number of edges that in the community that belong to both the CD44high and CD44low networks divided by the total number of edges in the community; an edge was deemed to belong to a community if it connected two nodes in the community. In order to estimate the probability of getting an equal or higher similarity value by chance, we first generated 1000 random configuration models with the same degree distribution as a given community, separately for the CD44high and CD44low derived networks. The Jaccard similarities of each randomly generated pair was then used to build a null distribution from which empirically estimate a p-value. The mean of this distribution was considered as the expected value.

Identification of pre-adapted cells. Two different strategies were employed to identify the pre-adapted cells. The first one takes advantage of SWNE; a threshold was applied on the first component and the cells showing extreme values (>0.75) were labelled as pre-adapted. The second strategy leverages random forests classifiers64. First of all, the data sets of CD44high cells in +E2 media and starved conditions (2 days) were split into training and testing sets, using 10% and 90% of the cells, respectively. The training set was then used to call the DEGs between the two conditions (+E2 vs starved), using the procedure described in the Differential expression analysis paragraph above. Those DEGs were used as input features to train a random forest classifier, using the randomForest R package (v4.6.14; default parameters). This model was then used to test the remaining data. Those cells in the testing set labelled as +E2 that were showing a probability >50% of being classified as starved were considered pre-adapted.

AUCell (R package v1.0.0) was used to quantify the activity of the pre-adapted signatures (and of other signatures, whenever indicated in the text) in single cells. First of all, normalised data were processed using the AUCell_buildRankings function. The resulting rankings, along with the signatures obtained, are then subject to a filtering step AUCel (calcAUCel function) using the top 5% (set to 5% of the number of input genes). Following inspection of the resulting distributions, thresholds were then manually set to 0.37, 0.18 and 0.32 for the signatures of pre-adapted cells either based on SWNE or random forests, or for the LTED signature (defined as those genes upregulated in LTED vs MCF7, as described in the Differential expression analysis section above).

Re-analysis of published primary samples. Bulk RNA-seq data sets for 1222 breast cancer samples were downloaded from the GDC (Genomic Data Commons) data portal (http://portal.gdc.cancer.gov) using gdc-client, according to metadata
obtained on July 25, 2018. Gene features were normalised to sequencing depth. Given that only a fraction of the samples was pre-classified using PAM5030, k-nearest neighbours (k-NN) classification was employed to impute the rest of the samples. This was performed via the knn function in the R package class (v7.3-14), using the pre-classified samples as the training data. Unclassified samples were ascribed to a particular subtype only when showing >60% probability of being assigned to that class. Spearman’s correlations between expression profiles of pairs of genes were calculated on the depth-normalised values. Prior to calculating signature scores, these numbers were further log2-transformed (pseudo-count set to 1) and scaled to z-score gene-wise.

Re-analysis of published profiles of CTCs. Normalised data for circulating tumour cells (CTCs) collected at five time points from a single patient along with identically processed blood specimens from 10 healthy donors10 were downloaded from GEO (GSE14245). For each capture, the log2-fold-change between EPCAM+ cells and the matched IgG+ cells (control) was calculated. DEGs were defined as those genes showing a linear fold change between EPCAM+ cells and control >= 1.5. The fraction of DEGs overlapping the genes in the pre-adapted signature was then calculated for each pair. To test if the observed difference between the fraction of DEGs in CTCs and in healthy specimens was random, a P-value was calculated using the Wilcoxon rank-sum test. The corresponding false discovery rate (FDR) was estimated by 1000 permutations. Raw data for individual CTC-clusters (median of three cells per cluster) and numerically matched pools of single CTCs from the same specimen36 were downloaded from GEO (GSE1827). Each profile was normalised by depth, then a profile-specific score was derived for the signature of the pre-adapted cells by summing the normalised expression values of all genes in the signature. These numbers were then divided by the maximum across all profiles. To test if the observed difference between the values obtained for the clusters against the matched pools of CTCs, a P-value was calculated using the Wilcoxon rank-sum test. The corresponding false discovery rate (FDR) was estimated by 1000 permutations.

Quantification of GFP-positive cells. A custom Python script (available upon request) was employed to segment images based on DAPI (to count the total number of cells) and GFP signal (to quantify the fraction of GFP+ cells).

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

Data availability
Raw sequencing data were deposited at the Gene Expression Omnibus (GEO) under accession number GSE12743. Processed data for single and clustered circulating tumour cells were obtained from the GEO (GSE14245 and GSE1827, respectively). Bulk RNA-sequencing datasets for luminal breast cancer were downloaded from the GDC (Genomic Data Commons)36 data portal (http://portal.gdc.cancer.gov; July 25, 2018). All the other data supporting the findings of this study are available within the article and its supplementary information files and from the corresponding authors upon reasonable request. A reporting summary for this article is available as a Supplementary Information file.

Code availability
The R scripts to reproduce the analyses and plots reported in this paper are available from the corresponding authors upon request.

Received: 26 November 2018 Accepted: 30 July 2019
Published online: 02 September 2019

References
1. Early Breast Cancer Trialists’ Collaborative Group (EBCTCG) et al. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. Lancet 378, 771–784 (2011).
2. Early Breast Cancer Trialists’ Collaborative Group (EBCTCG). Aromatase inhibitors versus tamoxifen in early breast cancer: patient-level meta-analysis of the randomised trials. Lancet 366, 1341–1352 (2015).
3. Asselain, B. et al. Long-term outcomes for neoadjuvant versus adjuvant chemotherapy in early breast cancer: meta-analysis of individual patient data from ten randomised trials. Lancet Oncol. 19, 27–39 (2018).
4. Pan, H. et al. 20-Year risks of breast-cancer recurrence after stopping endocrine therapy at 5 years. N. Engl. J. Med. 377, 1836–1846 (2017).
5. Cassaud, A. K. et al. Multiclonal invasion in breast tumors identified by topographic single cell sequencing. Cell 172, 205–217.e12 (2018).
6. Yap, T. A., Gerlinger, M., Futreal, P. A., Pusztai, L. & Swanton, C. Intratumor heterogeneity: seeing the wood for the trees. Sci. Transl. Med. 4, 127ps10 (2012).
7. Polyvak, K. Heterogeneity in breast cancer. J. Clin. Invest. 121, 3786–3788 (2011).
8. Turaljc, S. et al. Tracking cancer evolution reveals constrained routes to metastases: TRACERX renal cell. Cell 173, 581–594.e12 (2018).
9. Jalili-Hamani, M. et al. Tracking the evolution of non-small-cell lung cancer. N. Engl. J. Med. 376, 2109–2112 (2017).
10. Xue, Y. et al. An approach to suppress the evolution of resistance in BRAFV600E-mutant cancer. Nat. Med. 23, 929 (2017).
11. Nazarian, R. et al. Melanomas acquire resistance to B-RAF(V600E) inhibition by RTR or N-RAS upregulation. Nature 468, 973–977 (2010).
12. Misiale, S. et al. Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. Nature 486, 532–536 (2012).
13. Johnson, B. E. et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. Science 343, 189–193 (2014).
14. Robinson, D. R. et al. Activating ESR1 mutations in hormone-resistant metastatic breast cancer. Nat. Genet. 45, 1446–1451 (2013).
15. Magnani, L. et al. Acquired CYP19A1 amplification is an early specific mechanism of aromatase inhibitor resistance in ERα metastatic breast cancer. Nat. Genet. 49, 444–450 (2017).
16. Siegel, M. B. et al. Integrated RNA and DNA sequencing reveals early drivers of metastatic breast cancer. J. Clin. Invest. 128, 1371–1383 (2018).
17. Chatot, F. et al. Kinetics, prognostic and predictive values of ESR1 circulating mutations in metastatic breast cancer patients progressing on aromatase inhibitor. Oncotarget 7, 74448–74459 (2016).
18. Kim, C. et al. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. Cell 173, 879–893.e13 (2018).
19. Tirosch, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma into single-cell RNA-seq. Science 352, 189–196 (2016).
20. Hinohara, K. et al. KDM5 histone demethylase activity links cellular transcription to therapeutic resistance. Cancer Cell https://doi.org/10.1016/j.ccell.2018.10.014 (2018).
21. Patten, D. K. et al. Enhancer mapping uncovers phenotypic heterogeneity and evolution in patients with luminal breast cancer. Nat. Med. 24, 1469–1480 (2018).
22. Rambow, F. et al. Toward minimal residual disease-directed therapy in melanoma. Cell 174, 843–855.e19 (2018).
23. Ebinger, S. et al. Characterization of rare, dormant, and therapy-resistant cells in acute lymphoblastic leukemia. Cancer Cell 30, 849–862 (2016).
24. Sharma, S. V. et al. A Chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. Cell 141, 69–80 (2010).
25. Karayymaz, M. et al. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nat. Commun. 9, 3588 (2018).
26. Shaffer, S. M. et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature 546, 431–435 (2017).
27. Nguyen, Y. T. M. et al. Differential epigenetic reprogramming in response to specific endocrine therapies promotes cholesterol biosynthesis and cellular invasion. Nat. Commun. 6, 10045 (2015).
28. Martin, L.-A. et al. Discovery of naturally occurring ESR1 mutations in breast cancer cell lines modelling endocrine resistance. Nat. Commun. 8, 1865 (2017).
29. Martin, L.-A. et al. An in vitro model showing adaptation to long-term oestrogen deprivation highlights the clinical potential for targeting kinase pathways in combination with aromatase inhibition. Steroids 76, 772–776 (2011).
30. Pisco, A. O. et al. Non-Darwinian dynamics in therapy-induced cancer drug resistance. Nat. Commun. 4, 2467 (2013).
31. Razavi, P. et al. The genomic landscape of endocrine-resistant advanced breast cancers. Cancer Cell 34, 427–438.e6 (2018).
32. Wu, Y., Tamayo, P. & Zhang, K. Visualizing and interpreting single-cell gene expression datasets with similarity weighted nonnegative embedding. Cell Syst. 7, 656–666.e4 (2018).
33. Tang, D. G. Understanding cancer stem cell heterogeneity and plasticity. Cell Res. 22, 457–472 (2012).
34. Brooks, M. D., Burness, M. L. & Wicha, M. S. Therapeutic Implications of Cellular Heterogeneity and Plasticity in Breast Cancer. Cell Stem Cell 17, 260–271 (2015).
35. Chafer, C. L. et al. Poised chromatin at the ZEB1 promoter enables cancer cell plasticity and enhances tumorigenicity. Cell 154, 61–74 (2013).
36. Bausch-Fluck, D. et al. A mass spectrometric-derived cell surface protein atlas. PLoS ONE 10, e0121314 (2015).
37. Chen, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. 5, 251–267.e3 (2015).
38. Liberek, A. et al. The molecular signatures database hallmark gene set collection. Cell Syst. 1, 417–425 (2015).
39. Albar, S. et al. SCENIC: single-cell regulatory network inference and clustering. Nat. Methods 14, 1083–1086 (2017).
40. Grossman, R. L. et al. Toward a shared vision for cancer genomic data. N. Engl. J. Med. 375, 1109–1112 (2016).
41. Colleroni, M. et al. Annual hazard rates of recurrence for breast cancer during 24 years of follow-up: results from the international breast cancer study group trials I to V. J. Clin. Oncol. 34, 927–935 (2016).
42. Early Breast Cancer Trialists’ Collaborative Group (EBCTCG). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. Lancet 365, 1687–1717 (2005).
43. Perou, C. M. et al. The molecular signatures of breast cancer subtypes: a snapshot of 2018. Cell Syst. 1, e1000279 (2020).
44. Aceto, N. et al. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. Cell 158, 1110–1122 (2014).
45. Yu, M. et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. Science 339, 580–584 (2013).
46. Magnani, L. et al. SREBP1 drives KRT80-dependent cytосkeletal changes and invasive behavior in endocrine resistant ERα breast cancer. bioRxiv https://doi.org/10.1101/380634 (2018).
47. Lawson, D. A. et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. Nature 526, 131–135 (2015).
48. Puram, S. V. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell 171, 1611–1624. e24 (2017).
49. Torre, E. et al. Rare cell detection by single-cell RNA sequencing as guided by single-molecule RNA FISH. Cell Syst. 6, 171–179.e5 (2018).
50. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. 36, 411–420 (2018).
51. Lin, Y. et al. Evaluating stably expressed genes in single cells. bioRxiv https://doi.org/10.1101/229013 (2018).
52. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. F1000Research 5, 2122 (2016).
53. Patel, A. P. et al. Single-cell RNA-seq highlights intratumat heterogeneity in primary glioblastoma. Science 344, 1396–1401 (2014).
54. Nguyen, Q. H. et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. Nat. Commun. 9, 2028 (2018).
55. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. 4, 1184–1191 (2009).
56. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
57. Li, H. et al. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
58. Tatevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. PLoS Comput. Biol. 12, e1004873 (2016).
59. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 16, 278 (2015).
60. Benjamini, Y., Hochberg, Y. & Benjaminit, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Royal Statist. Soc. Series B (Methodological) 57, 289–300 (1995).
61. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. Bioinformatics 21, 3940–3941 (2005).
62. Breiman, L. Random forests. Mach. Learn 45, 5–32 (2001).
63. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. 27, 1160–1167 (2009).

Acknowledgements

We want to acknowledge and thank all patients and their families for the support and for donating the research samples. The authors gratefully acknowledge infrastructure support from the Cancer Research UK Imperial Centre, the Imperial Experimental Cancer Medicine Centre and the National Institute for Health Research Imperial Biomedical Research Centre. L.M. was supported by a CRUK fellowship (C46704/A23110) and an Imperial Junior Fellowship (G53019). S.P.H. was supported by a Basic Science Research Program through the National Research Foundation of Korea (NRF-2013R1A1A1011832) and by a CRUK programme award (CRUK/C37/A18784). I.B. was supported by CRUK funding (C46704/A23110) and by an Imperial College Research Fellowship. G.C. was supported by a Marie Skłodowska Curie Training Grant (642691, EpiPredict). The Imperial College Healthcare NHS Trust Tissue Bank provided tissue samples. Consent was collected at IEO and Imperial College (Project R15036) by the respective Tissue Banks. Other investigators may have received samples from these same tissues. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Author contributions

S.P.H., I.B. and L.M. planned the research; S.P.H. and Y.L. performed the experiments; I.B., S.P.H., T.E.C., K.R.M. and L.M. analysed the data; G.C., N.R., G.P., S.B., A.R. and R.C.C. contributed technical support; I.B., S.P.H. and L.M. wrote the paper.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-11721-9.

Competing interests: One of the authors, Y.L., is an editor on the staff of Nature Communications, but was not in any way involved in the journal review process. All the other authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Peer review information: Nature Communications thanks Tallulah Andrews and other anonymous reviewer(s) for their contribution to the peer review of this work.

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