nature communications

Article

https://doi.org/10.1038/s41467-024-51424-4

Multi-omic lineage tracing predicts the transcriptional, epigenetic and genetic determinants of cancer evolution

Received: 10 August 2023

Accepted: 5 August 2024

Published online: 01 September 2024

Check for updates

F. Nadalin[1,2,5] ✉, M. J. Marzi[1,5], M. Pirra Piscazzi[1,5], P. Fuentes-Bravo[1,5], S. Procaccia[1], M. Climent[1], P. Bonetti[1], C. Rubolino[1], B. Giuliani[1], I. Papatheodorou[2], J. C. Marioni[2,3,4] & F. Nicassio[1] ✉

Cancer is a highly heterogeneous disease, where phenotypically distinct subpopulations coexist and can be primed to different fates. Both genetic and epigenetic factors may drive cancer evolution, however little is known about whether and how such a process is pre-encoded in cancer clones. Using single-cell multi-omic lineage tracing and phenotypic assays, we investigate the predictive features of either tumour initiation or drug tolerance within the same cancer population. Clones primed to tumour initiation in vivo display two distinct transcriptional states at baseline. Remarkably, these states share a distinctive DNA accessibility profile, highlighting an epigenetic basis for tumour initiation. The drug tolerant niche is also largely pre-encoded, but only partially overlaps the tumour-initiating one and evolves following two genetically and transcriptionally distinct trajectories. Our study highlights coexisting genetic, epigenetic and transcriptional determinants of cancer evolution, unravelling the molecular complexity of pre-encoded tumour phenotypes.

Cancer adopts evolutionary pathways that sustain the disease. Aggressive tumour behaviours, such as the dissemination to distant organs, diminished susceptibility to treatment, and disease relapse, result from either selection or adaptation processes, possibly intertwined[1]. When a selective process occurs, the fate of a cancer clone is determined at the root of the evolutionary process. In this case, the heterogeneity of tumour phenotypes can, at least in principle, be identified ahead of selection[2]. The pre-existence of aggressive phenotypes has been linked to the so-called cancer stem cell (CSC) theory[3] and observed in leukaemia[4,5] and solid tumours, such as colon[6] and breast cancer[7,8], as well as glioma[9,10]. According to such a model, tumour cells are not all equal, instead a stem-like cancer niche exists that is primed to sustain most of the aggressive phenotypes, such as tumour re-initiation, metastatic dissemination potential, and capacity to survive cytotoxic treatments[11].

Predicting cancer phenotypes requires linking the molecular state of a clone to its fate with high precision. Without a priori information, tumour phylogeny can be inferred from somatic mutations[12–15]; however, this approach is limited by the high sparsity of single-cell data. Single-cell lineage tracing consists in inserting barcodes in the genome of the cells with the aim of tracing their progeny[16–19]. In cancer, this approach has been used to investigate clonality in metastases[20], survival upon cytotoxic treatment[21,22], as well as to dissect the clonal origin of the primary tumour and metastasis growth[23–26], possibly in vivo[27]. However, these studies mainly focus on the evolutionary trajectories, rather than on the driving molecular features of pre-existing phenotypes.

[1]Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia (IIT), Milan, Italy. [2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK. [3]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [4]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK. [5]These authors contributed equally: F. Nadalin, M. J. Marzi, M. Pirra Piscazzi, P. Fuentes-Bravo. ✉e-mail: francesca@ebi.ac.uk; francesco.nicassio@iit.it

Tumour evolutionary diversity can have either a genetic or non-genetic origin[28,29]. Single-cell multi-omics has recently emerged as a promising tool to study cancer evolution[30]. Here, we combine single-cell multi-omics with lineage tracing in a unique framework, which allows simultaneous clonal, gene expression, and chromatin accessibility profiling at single-cell resolution. Using phenotypic assays on barcoded cells, we identify the clones endowed with aggressive cancer behaviours typical of the stem-like cancer niche, specifically tumour-initiating capacity and drug tolerance. Subsequently, we extract robust transcriptional, epigenetic, and genetic features of naïve cells and associate them to clonal subpopulations. By integrating these multiple layers of information, we identify the regulatory elements that predict cancer evolution in response to adverse environmental conditions. Finally, by tracing the transcriptional changes of clones across time, we unravel the role of pre-existing molecular features in shaping the differentiation breadth of stem-like subpopulations.

Results

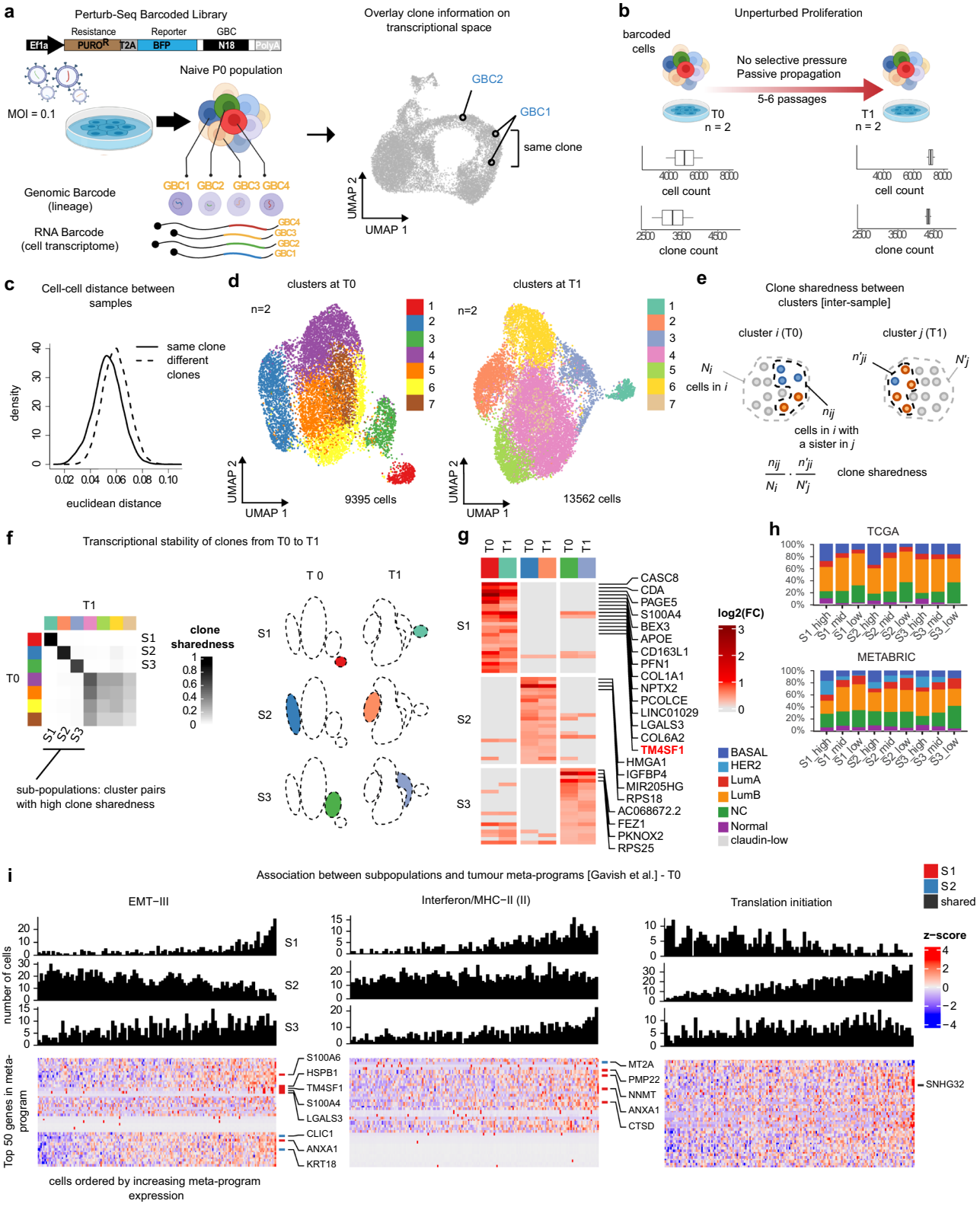
SUM159PT exhibits high transcriptional plasticity and comprises three distinct subpopulations: S1, S2, and S3

To investigate the potential of cancer cells to promote tumour initiation and escape cytotoxic treatment, we combined single-cell sequencing with phenotypic assays. We selected SUM159PT, a triple-negative breast cancer cell line (TNBC), as a model system. SUM159PT belongs to the claudin-low mesenchymal subtype[31] and is characterised by (i) a nearly diploid genotype, but bearing a specific set of mutations typically associated with TNBC (*HRAS*, *PIK3CA*, *TP53* and *MYC* amplification[32]); (ii) an intrinsic heterogeneity, with an underlying variability in the expression of epithelial and mesenchymal genes and a small proportion of cells in a CSC state[33,34]; and (iii) an aggressive phenotype driven by the CSC component, which is highly tumourigenic and invasive in vivo[33–35]. A single-cell lineage tracing approach was used to link the molecular state of a cell to its fate (Fig. 1a and Supplementary Fig. 1a). To obtain ~10,000 distinct genetic barcodes (GBC), 100,000 SUM159PT cells were infected with a lentiviral pool at a multiplicity of infection (MOI) = 0.1 and subsequently FAC-sorted to retain only the transduced fraction[18]. Endogenous as well as GBC-carrying transcripts were then captured by single-cell RNA-seq (scRNA-seq). The parental population was sampled and processed at two time points, T0 and T1, separated by 13–15 days (Fig. 1b). At basal state, between 5017 and 5996 unique clones were found in the two replicates and between 83% and 88% of high-quality cells were assigned a clone identity (Supplementary Fig. 1b and Supplementary Data 1), making the lineage of even rare cell subpopulations accessible to analysis. The distribution of clones at the two time points was similar, highlighting that no spontaneous clone selection occurs in the timeframe (Supplementary Fig. 1c). Moreover, 68% and 57% of the clones respectively detected in T0 and T1 were shared between the two time points, with >50% of clones constituted by a single cell in each sample (Supplementary Fig. 1d and Supplementary Data 1). When evaluating the relationship between clonality and gene expression profile at basal state, cells stemming from a common clone at the moment of infection, hereafter *sister cells*, were on average only slightly more similar to one another compared to non-sisters (Fig. 1c). We next asked whether the transcriptional similarity between sister cells is clone-specific—in other words, whether some clones show a distinctive gene expression profile and other clones are more plastic [a similar approach is proposed in ref. 17]. We detected seven distinct gene expression clusters in T0 and T1, respectively (Fig. 1d, Supplementary Fig. 1e, f and Supplementary Data 1) and compared the clone content of every cluster pair across the two time points using a *clone sharedness* score (see section "Methods" and Fig. 1e). Most clones that clustered together in T0 were mapped to multiple distinct clusters in T1, and vice versa, suggesting a high transcriptional plasticity already at baseline. In contrast, three cluster pairs in T0 and T1, respectively, comprising 28% and 23% of the cells at the two time points, showed mutually high clone sharedness (see section "Methods" and Fig. 1f); we conclude that these subpopulations are transcriptionally stable and we will refer to them as S1, S2, and S3 hereafter. They respectively comprise 3.6%, 14.7%, and 7.4% cells on average. We obtained a gene expression signature for each of them (Fig. 1g and Supplementary Data 2) that is independent of cell-culture effect. Of note, S1 was enriched in genes involved in collagen processing and matrix remodelling (Supplementary Fig. 1g and Supplementary Data 3). S1 cells showed upregulation of *S100A4*, a gene associated with metastatic behaviours[20,36,37], and *TM4SF1*, whose role in promoting cell proliferation and invasion in epithelial tumours has been assessed[38–41]. The microRNA-205 host gene (*MIR205HG*) was found as S2-specific and has been associated to basal cells[42], epithelial-to-mesenchymal (EMT) transition, and multiple cancer diseases[43,44]. The oncogene *HMGA1* is part of the S2 signature and has been associated to the TNBC subtype[45]. S3 was distinguished by the expression of *FEZ1*, a microtubule adaptor[46], and *RPS25*, a gene acting on cellular response to stress by downregulating p53[47]. In conclusion, single-cell lineage tracing revealed that SUM159PT exhibits high transcriptional plasticity, but comprises three distinct, transcriptionally stable subpopulations.

SUM159PT transcriptional heterogeneity is recapitulated in primary tumours

To assess the relevance of stable SUM159PT transcriptional programmes in primary TNBC tumours, we leveraged the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC[48]) and The Cancer Genome Atlas (TCGA[49]). Figure 1h reports the stratification of breast cancer tumours according to high, medium and low gene expression classes (see section "Methods") in TCGA and METABRIC datasets. S1 and S2 signatures were associated with the basal tumour subtype (including claudin-low, which accounts for 9.8% of all tumours in METABRIC) in both datasets (adj. p value < 0.001), whereas S1 and S3 with the claudin-low subtype (adj. p value < 0.001; Benjamini–Hochberg correction), suggesting that the stable transcriptional programmes identified in SUM159PT capture broad basal tumour features. We noted that the S1 signature genes were organised into few distinct co-expression blocks (Supplementary Fig. 2a), hinting that they may be part of a network also in tumours. Therefore, we next evaluated whether S1, S2, and S3 recapitulate intra-tumour heterogeneity in scRNA-seq datasets from primary samples. In primary TNBC tumours[50], we could detect both S1 and S3 programmes and identify $S1^+$ and $S3^+$ cell subsets accordingly (Supplementary Fig. 3a–c); in particular, we detected a strong upregulation of *S100A4* in the $S1^+$ subset (see Supplementary Fig. 3d and Supplementary Data 4). Then, we leveraged the Curated Cancer Cell Atlas (3CA, https://www.weizmann.ac.il/sites/3CA/), containing scRNA-seq data for over 1000 primary tumour samples from over 70 studies, along with the associated gene meta-programmes[51]. We reasoned that a high association of SUM159PT subpopulations with key tumour meta-programmes would be a strong indication for the generalisability of the signatures we defined. We noted that the aggregate meta-programme expression across SUM159PT clusters was non-stochastic (Supplementary Fig. 2b), suggesting a common pattern of gene expression heterogeneity; specifically, the 3CA meta-programme "EMT-III" was enriched in S1 cells, "Interferon/MHC-II (II)" both in S1 and S3 cells, and "Translation initiation" in S2 cells (Fig. 1i and Supplementary Fig. 2c), in agreement with pathway enrichment analysis (see Supplementary Fig. 1f). "EMT-III" contains genes involved in the maintenance of a hybrid EMT state and belonging to the S1 signature (e.g., *S100A4*, *TM4SF1*, and *LGALS3*); this meta-programme is recurrent across donors and cancer types, notably in breast[51]. Finally, we performed scRNA-seq on the TNBC cell line MDA-MB-231 TGL. Two of the seven clusters we detected in MDA-MB-231 TGL showed high expression of top S1 signature genes (*CDA*, *S100A4*, *LGALS3*, *COL6A1*, *COL6A2*;

https://doi.org/10.1038/s41467-024-51424-4



Supplementary Fig. 2d, e and Supplementary Data 5, 6); importantly, these clusters also showed high "EMT-III" meta-programme expression (Supplementary Fig. 2f, g). Taken together, these results suggest that the transcriptionally stable signatures of SUM159PT, notably S1, are recurrent in other TNBC models and primary breast tumours, and are also shared across other cancer types.

Cancer clones promote tumour initiation in a non-stochastic manner

To investigate the tumour-initiating capacity of SUM159PT, we transplanted barcode-labelled cells into the mammary fat pads of nine NSG (NOD/SCID/IL2Rγ$_c^{-/-}$) immunodeficient mice and then evaluated the barcode composition in each primary tumour. We isolated tumour cells

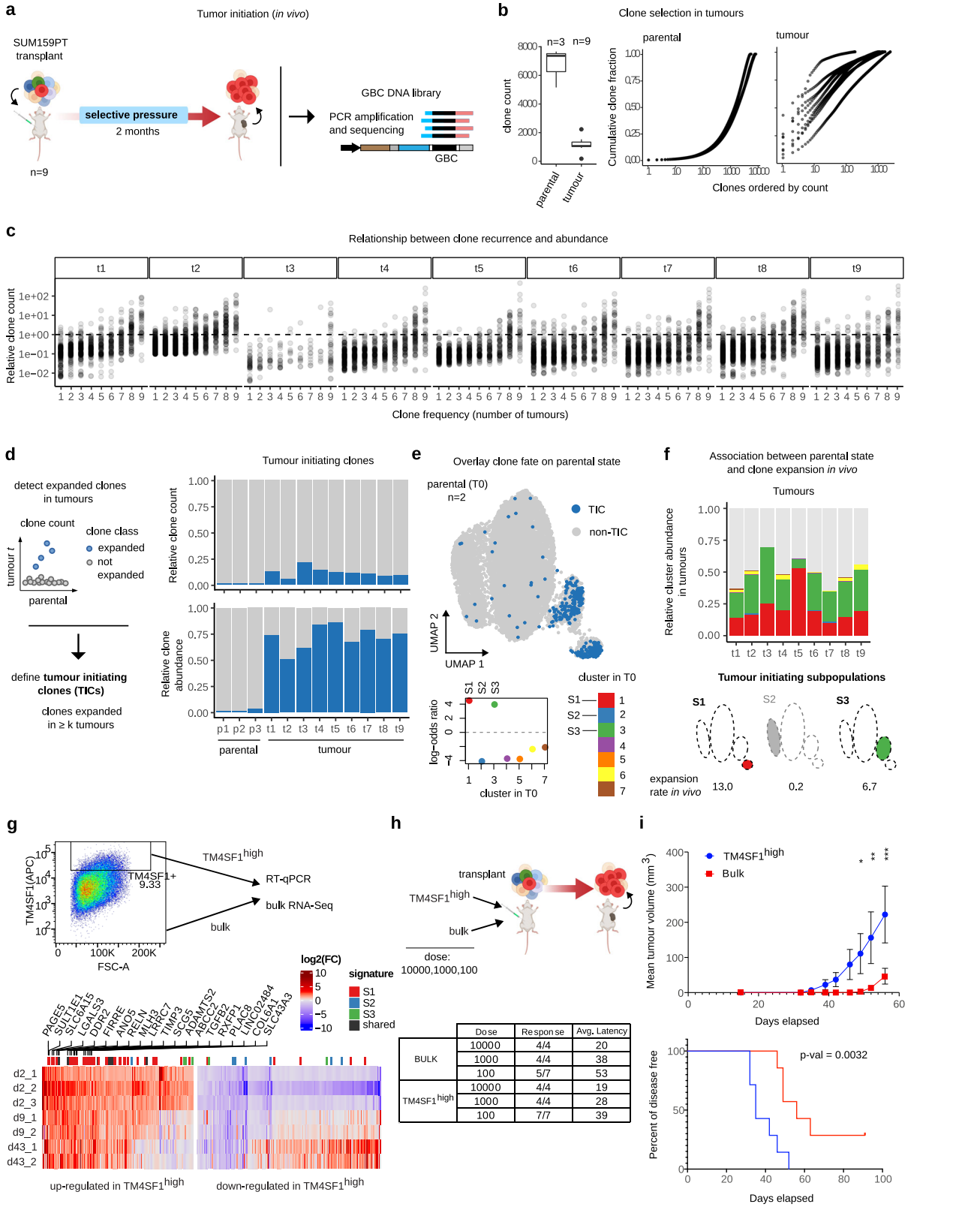
Fig. 1 | Lineage tracing identifies transcriptionally stable TNBC cell subpopulations. a SUM159PT cells were infected with a lentiviral library of unique barcodes (Perturb-seq GBC library) at 0.1 multiplicity of infection (MOI). The readout of each cell is its lineage (genetic barcodes) and gene expression profile (3′-end cDNA sequencing). Clone information is overlaid on single-cell gene expression space. **b** Experimental design, passive propagation. Top: barcoded SUM159PT cells from the same infection experiment were processed by scRNA-seq at two passages (T0 and T1, $n = 2$ replicates each). Bottom: number of detected clones and cells assigned to clones. **c** Gaussian kernel density of Euclidean distances between sister cells (solid line) and non-sister cells (dashed line) computed on a joint T0 and T1 space (see section "Methods"). **d** UMAP representation of T0 (9395 cells) and T1 (13,562 cells) coloured by cluster; only cells assigned to clones are shown. **e** Definition of clone sharedness score between clusters *i* and *j*. **f** Left: rows are clusters in T0 (as in **d**), columns are clusters in T1, and entries are clone sharedness scores for each pair. Rows and columns are ordered by higher to lower scores. The three top-scoring pairs, referred to as subpopulations (S1, S2, S3), are shown on UMAP (right). **g** Subpopulation gene signatures. Rows are the 25 genes in the subpopulation signature showing the highest $\log_2$(FC) in T0, columns are subpopulations split by time point, and entries are $\log_2$(FC) values between a subpopulation and its complement at the same time point. The top 15 (S1) and top 4 (S2, S3) genes are labelled. The surface marker TM4SF1, highlighted in red, is used for sorting the S1 subpopulation. **h** Stratification of breast cancer samples into molecular subtypes by subpopulation signature activity in TCGA (top) and METABRIC (bottom) datasets (NC not classified). **i** Association with tumour meta-programmes from the Curated Cancer Cell Atlas. The columns are the cells at T0 ordered by non-decreasing module score, computed on the union of the top 50 significant genes of the meta-programme (in rows); genes in S1 and S2 signatures are labelled. The bar plots show the binned cell count for each subpopulation [**a**, **b** created with Biorender.com released under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license].

and extracted the genomic DNA (gDNA), which was then amplified and sequenced (Fig. 2a). Noteworthy, the GBC count measured in bulk (parental cells) recapitulates the actual clone abundance, measured as the relative number of cells per clone in single-cell samples (Supplementary Fig. 4a, b). Clone selection was heterogeneous across tumours −a stochastic effect that we observed also in other cohorts of tumours transplanted in mice (data not shown). Only between 3% and 33% of SUM159PT clones contributed to tumour formation (Fig. 2c and Supplementary Data 7), showing a deep clone selection. The size of clonal subpopulations greatly varied within a tumour; on average, the top 1% abundant clones covered more than 50% of the entire tumour mass (Supplementary Data 7), and this was not merely a consequence of higher initial abundance (see below). This picture suggests a variable tumour-initiation potential among surviving clones in vivo. In epithelial cancers, the tumour-initiation potential has been regarded as an intrinsic feature of cells, rather than a feature acquired during tumour formation[8]. Consistently, we observed that a limited set of clones was recurrent and covered a high proportion of the tumour mass compared to sporadic ones (Fig. 2c). To exclude bias in the detection of expanded clones, we considered GBC abundance relative to pre-implantation (Fisher's exact test, see section "Methods"). In total, 138 clones were significantly more abundant in at least 6 out of 9 tumours compared to their average abundance at basal state. We refer to these as *tumour-initiating clones* (TICs) hereafter (Fig. 2d). We conclude that the tumour-initiating capacity of SUM159PT cells is largely pre-encoded.

The baseline programmes S1 and S3 predict tumour initiation

To determine which transcriptional states are primed to tumour initiation, we traced TICs back to their parental population. TICs were strongly associated with S1 and S3 transcriptomes at baseline, with the two subpopulations showing a similarly strong enrichment (average odds ratio 4.4 and 4.1 for S1 and S3, respectively; Fig. 2e, Supplementary Fig. 4c and Supplementary Data 7). Both S1 and S3 were transcriptionally stable in culture in a timeframe of 2 weeks, as shown in Fig. 1f, suggesting that the gene expression profile of TICs at baseline may be predictive of the phenotype. S1 and S3 gene signatures partially overlap: 9 out of the 29 gene expression markers of S3 are shared with S1, including some of the top ranked in S1 (*COL1A1*, *NPTX2*, *NREP*; see Fig. 1g). However, the two subpopulations showed different gene expression patterns, with S1 being clearly separated from all the other clusters in gene expression space (average silhouette width 0.29 and 0.30, respectively; Supplementary Fig. 1e), suggesting that TICs stem from two distinct transcriptional programmes. We then asked whether the clones in S1 and S3 differ in terms of their expansion potential in vivo. When the subpopulation identity was mapped on tumours, clones in S1 and S3 highly contributed to the tumour mass, relative to their initial abundance at baseline, compared to the other clones (Fig. 2f and Supplementary Fig. 4d); upon transplant, the expansion rate of S1 was high compared to S3 (12.7-fold for S1 and 6.9-fold for S3, on average; Supplementary Data 7). We profiled the transcriptome of SUM159PT tumours by bulk RNA-Seq. Neither the S1 nor the S3 signatures as a whole were upregulated in SUM159PT tumours with respect to the parental population, including some among the top significant genes (Supplementary Fig. 4e, f and Supplementary Data 8), hinting that clones undergo transcriptional reprogramming upon transplant. This change in gene expression profile is in line with the cancer stem-cell hypothesis, where a small, stem-like, cell subpopulation exhibits both tumourigenic and differentiation potential. Of note, *S100A4*, *TM4SF1*, and *LGALS3*, three among the top significant genes in the S1 signature belonging to the EMT-III meta-programme (see Fig. 1g and Supplementary Fig. 2c), were upregulated in SUM159PT tumours ($\log_2$(FC) = 2.24, 3.40, 3.76 and adj. *p* value = 1.09e − 27, 2.85e − 34, 2.09e − 30, respectively). TIC signature gene expression was persistent in metastatic pancreatic cancer mouse models[20] and in the pre-metastatic niche of lung adenocarcinoma[52] (Supplementary Fig. 4g, h). Notably, the metastatic potential in these tumour models has been associated with the adoption of late hybrid EMT states and the activation RUNX2, a transcription factor mediating extracellular matrix remodelling, in agreement with our findings (see Fig. 1i and Supplementary Figs. 1g, 2g). We conclude that the tumour-initiating niche of SUM159PT shares markers across different cancer diseases and, although plastic, could be partially reminiscent of its molecular state at baseline. To directly verify the tumour-initiating potential of TICs, we searched for surface markers for prospective isolation and identified transmembrane 4 L6 family member 1 (TM4SF1), 1 of the top 20 significant genes of the S1 signature and highly upregulated in SUM159PT tumours; we could not identify any S3-specific surface marker. High TM4SF1 protein expression has been linked to CSCs and previously employed for prospective isolation of cancer subpopulations in human and murine breast models[53,54]. Therefore, we set up a strategy for isolating TM4SF1^{high} cells by FAC-sorting (gated on top 5%; Fig. 2g and Supplementary Fig. 5a–d); the TM4SF1^{high} population showed extensive upregulation of several genes in the S1 signature compared to the bulk population, and this was not the case for genes in S2 and S3 signatures (RT-qPCR and RNA-seq; Fig. 2g and Supplementary Fig. 6a–c). Of note, the expression of the S1 signature was maintained even after several passages in culture (Supplementary Fig. 6c and Supplementary Data 9). TM4SF1^{high}-associated genes were mainly related to invasion and metastasis pathways and suggestive of TWIST1, STAT3, and HIF1A activation (Supplementary Fig. 6d). Limiting dilution transplantation is a well-established approach to quantify the tumour-initiating content of a cell population. We injected orthotopically serial dilutions from bulk and TM4SF1^{high} populations (Fig. 2h and Supplementary Fig. 5a) into NSG (NOD/SCID/IL2Rγ$_c^{-/-}$) immunodeficient mice. At the lowest dilution (100 cells), TIC number is a limiting factor and TM4SF1^{high} cells developed tumours with higher efficiency than mice transplanted with the same number of bulk cells (26% average latency reduction; Fig. 2i and Supplementary Fig. 6e, g), suggesting that the TIC content of the TM4SF1^{high} subpopulation is higher. We conclude that S1 holds an

https://doi.org/10.1038/s41467-024-51424-4



	Dose	Response	Avg. Latency
BULK	10000	4/4	20
	1000	4/4	38
	100	5/7	53
TM4SF1high	10000	4/4	19
	1000	4/4	28
	100	7/7	39

increased tumour-initiating capacity compared to the whole SUM159PT population.

The S3 programme confers a selective growth advantage upon chemotherapy

We next investigated the response of cancer clones upon drug response in vitro on cultured cells and in vivo on transplanted tumours, using paclitaxel, an anti-mitotic chemotherapy agent used to treat many cancer types[55]. We treated barcode-labelled SUM159PT cells at 50 nM (which corresponds to ~IC95; Supplementary Fig. 7a) for 3 days in culture, with the untreated condition as a control (Fig. 3a, top). To evaluate the drug response in vivo, we transplanted barcode-labelled SUM159PT cells into the mammary fat pads of six NSG immunodeficient mice; once tumour was formed, mice

Fig. 2 | Tumour-initiating clones are recurrent and originate from S1 and S3 subpopulations. a Tumour-initiation assay. Barcoded SUM159PT cells were injected orthotopically in NSG mice; gDNA from parental ($n = 3$) and tumours ($n = 9$) were sequenced. **b** Left: clone count as number of distinct GBCs (bounds of box: upper (q75) and lower (q25) quartiles; centre: median; upper whisker: min{max(x), q75 + 1.5·IQR}; lower whisker: max{min(x), q25 − 1.5·IQR}). Right: cumulative clone frequency, where GBCs are ordered by non-increasing abundance. **c** Each graph refers to a tumour and each dot is a clone; clones are grouped by the number of times they are observed across tumours ($x = k$) and their frequency over the total tumour size is shown on y. **d** Left: detection of tumour-initiating clones (TIC) by comparison between clone abundance in tumour t in the parental population. Right: fraction of clones (top) and relative clone abundance (bottom), in parental and tumour samples, grouped by clone class. **e** Mapping of TICs at parental state (T0). Top: UMAP representation of T0 cells in gene expression space (TICs in blue). Bottom: log-odds ratio of cluster assignment vs TIC labelling at T0. **f** Association between T0 clusters and clone expansion in vivo. Top: normalised cluster abundance in each tumour (unassigned clones in grey). Bottom: TIC odds ratio in subpopulations. **g** Prospective isolation of S1 cells by FAC-sorting with TM4SF1 antibody (see Fig. 1g). Top: gate used for TM4SF1^{high} sorting. Bottom: differentially expressed genes (RNA-seq) between TM4SF1^{high} and bulk at days 0, 9, and 43 ($n = 2$ each). Entries are expression $\log_2$(FC) between conditions at the same time point. scRNA-seq and Multiome gene signatures are highlighted in colour (shared genes in black) and the 20 top upregulated genes in TM4SF1^{high} cells labelled. **h** TM4SF1^{high} cells are enriched for TICs. Top: TM4SF1^{high} or bulk cells are injected orthotopically at different dilutions. Bottom: response and average latency. **i** Tumour growth and disease-free survival. Top: growth dynamics (in days) of each primary tumour derived from transplantation of 100 cells ($n = 7$). Data are mean ± SEM. Asterisks mark the significance two-sided, unpaired t-test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.01$). Bottom: Kaplan–Meier curve reporting the time-dependent appearance of primary tumours derived from injection of 100 cells (Log-rank Mantel–Cox test) [**a**, **h** Created with Biorender.com released under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license].

were treated with paclitaxel every 5 days (Fig. 3a, bottom). In vitro, treatment induced a deep clonal selection: between 9% and 22% of the initial clone pool survived ≥10 days post-paclitaxel removal, while cells cultured for a comparable time span in the absence of treatment did not undergo clonal selection (Fig. 3b, top, and Supplementary Fig. 7c and Supplementary Data 7). A comparable effect was observed in an independent barcoding experiment (Supplementary Fig. 7c). In vivo, paclitaxel treatment delayed tumour growth, but did not trigger remission of the disease (Supplementary Fig. 7d); the major driver of clone selection was the tumour-initiation capacity (Fig. 3b, bottom). Of note, the clones able to survive and expand were not randomly selected, but recurrent upon independent treatments, both in vitro and in vivo (Supplementary Fig. 7e, f); therefore, we reasoned that both survival and proliferation potential were pre-encoded. We defined the *drug-tolerant clone* (DTC) pool as the set of clones that were significantly more abundant after treatment in at least four out of six samples compared to their average abundance at basal state (Fisher's exact test; see section "Methods" and Fig. 3c). We detected 171 and 164 DTCs in vitro and in vivo, respectively. When traced back to their baseline transcriptional state, clones surviving drug insult in vitro were depleted in S1, but were more abundant in S3 than expected by chance (Fig. 3d, left, and Supplementary Fig. 8a, c and Supplementary Data 1). In contrast, clones surviving drug treatment in vivo were belonging to either S1 or S3 (Fig. 3d, right, and Supplementary Fig. 8b and Supplementary Data 1). Note that 71% of TICs were also drug-tolerant in vivo (Supplementary Fig. 8d and Supplementary Data 7), confirming that the effect of paclitaxel in vivo was modest. When assessing the relative abundance of S1 and S3 clones in tumours treated with paclitaxel, S3 showed a higher fitness over S1 (Fig. 3e, and Supplementary Fig. 8e and Supplementary Data 7), in agreement with in vitro results. We deduced that the drug tolerance phenotype is different from the tumour-initiating capacity in SUM159PT and is primarily associated with the S3 baseline programme.

GALILEO links cancer clones with transcriptional programmes and DNA accessibility states at single-cell resolution

To investigate the epigenetic state of cancer clones and relate it to the transcriptional readout, we developed Genomic bArcoding pLus sIngLE-cell multi-Omics (GALILEO). Specifically, we performed single-cell Multiome ATAC plus gene expression sequencing on barcode-labelled SUM159PT nuclei in two biological replicates (Fig. 4a). This assay enables access to gene expression, DNA accessibility, and clone information simultaneously at single-nucleus resolution. At baseline, we identified 2023–2024 unique clones and assigned a clonal identity to between 73% and 86% of high-quality nuclei (Supplementary Fig. 9a and Supplementary Data 1). We obtained seven and six clusters for the two replicates, respectively, and retrieved the subpopulations S1, S2, and S3 previously identified by scRNA-seq (Fig. 4b and Supplementary Fig. 9b, c), with largely overlapping gene signatures (Supplementary Fig. 9d and Supplementary Data 10). When evaluating their DNA accessibility state, nuclei from S1, S2, and S3 showed distinct profiles in the scATAC-seq space, comparable to the scRNA-seq results (Supplementary Fig. 9e). To identify patterns of co-accessibility in the set of ~10^5 regions detected from scATAC-seq, we used cisTopic[56], a tool for scATAC-seq data analysis based on a topic modelling framework. Briefly, *topics*[56] are hidden variables represented as probability values across all ATAC regions in the dataset, and, conversely, cells are represented as probability values over topics; the benefit of this approach is that the number of topics is typically much smaller than the number of regions. Groups of regions and groups of cells where these regions are co-accessible are thus simultaneously captured via their association with topics. Next, we compared the region probability across every topic pair between the two replicates using the irreproducible discovery rate (IDR; see section "Methods" and Supplementary Fig. 10a). We defined a subset of reproducible regions (i.e., satisfying IDR < 0.05) for each topic pair, referred to as *ATAC module* hereafter, and assigned a *reproducibility score* to them (Fig. 4c, left, and Supplementary Fig. 10b and Supplementary Data 11). Note that our approach based on topic modelling discards ubiquitously accessible regions; combined with IDR filtering, this results in a substantial reduction of the size of the dataset (see pie chart in Fig. 4c). Most reproducible regions were found in few, large ATAC modules containing more than 400 regions, the largest one containing 1511 regions (Fig. 4d and Supplementary Fig. 10c). This few-to-few mapping across replicates suggests that the grouping of the regions into ATAC modules is non-stochastic. Therefore, each of these modules is expected to identify a pool of genomic elements that jointly participate in the regulation of gene expression. More than 90% of the regions could be assigned a regulatory element, according to the ENCODE cCRE registry: 9% were annotated as promoter-like signatures (PLS), within 200 bp of transcription start sites (TSS) of genes, 7% and 75% as proximal and distal enhancer-like signatures (pELS, dELS), respectively (Fig. 4c).

ATAC modules recapitulate the multiple DNA accessibility profiles of gene expression clusters

The regions found in the same ATAC module are accessible in the same cells, and, by definition, are reproducible across replicates. We investigated the relationship between ATAC modules and transcriptional or clonal subpopulations. We computed a score for each ATAC module M and for each cell c (AUC; see section "Methods"), representing the overall accessibility of the regions of M in c, and referred to as *module AUC* hereafter. Using module AUC, we associated several large ATAC modules with either S1, S2, or S3 (Fig. 4d and Supplementary Fig. 10c). Among the 20 highly reproducible modules, 8 (highlighted in bold in Fig. 4d) could be associated to any of S1, S2, or S3, with Module 1 (1511 regions) being the top predictor of both S1 (AUC = 0.98) and S3

https://doi.org/10.1038/s41467-024-51424-4

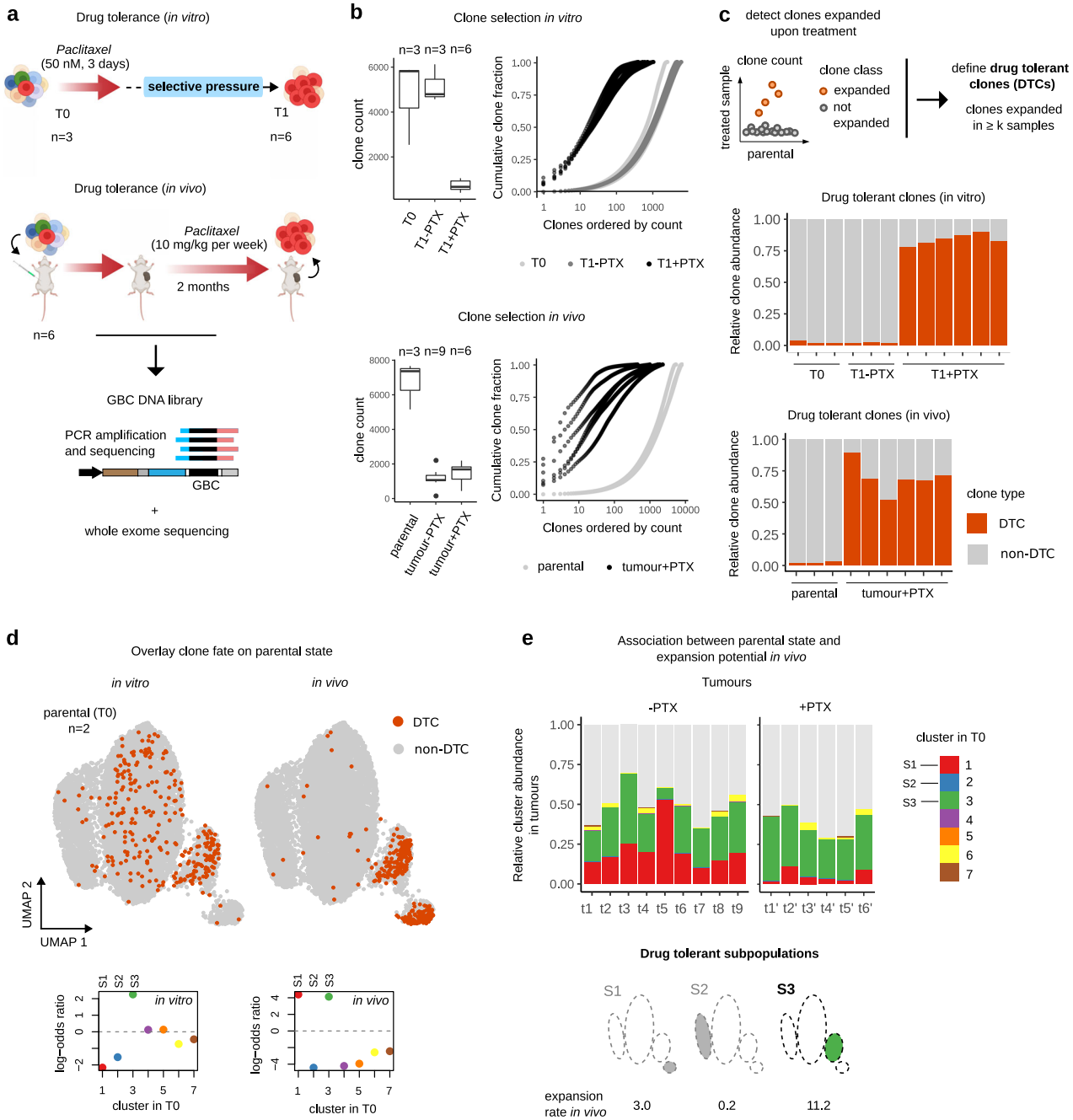


Fig. 3 | Drug-tolerant clones are recurrent and are enriched in S3 cells. **a** Experimental design, drug tolerance assay. Top: in vitro assay. Barcoded SUM159PT cells were treated with paclitaxel in vitro and harvested when single-cell colonies were grown ($n = 6$). GBC loci were PCR amplified and sequenced. The untreated parental population at T0 ($n = 3$) and T1 ($n = 3$) was also sequenced as a control. Bottom: in vivo assay. Barcoded SUM159PT cells were injected orthotopically in NSG mice; after tumour formation, mice were treated or not (see Fig. 2a) with paclitaxel. Parental samples ($n = 3$) were also sequenced as a control. **b** Clone selection upon treatment. Top: comparison of total clone count (left) and cumulative clone distribution (right) in parental, untreated, and treated in vitro samples (bounds of box: upper (q75) and lower (q25) quartiles; centre: median; upper whisker: min{max(x), q75 + 1.5·IQR}; lower whisker: max{min(x), q25 − 1.5·IQR}). Bottom: same as above, for parental, untreated tumour (as in Fig. 2a), and treated tumour samples (see also Fig. 2b legend). **c** Detection of drug-tolerant clones (DTC). Top: cartoon showing the comparison between clone abundance in each sample compared to the (average) abundance in the parental population; clones significantly more abundant in $k = 4$ out of 6 samples are defined as drug tolerant. Bottom: bar plot showing the relative clone abundance in parental and treated samples in vitro and in vivo, respectively, and grouped by class (drug tolerant or not). **d** Mapping of the drug-tolerant clones at parental state (T0). Top: UMAP representation of T0 cells on gene expression space, with cells classified as DTC in vitro (left) or in vivo (right) coloured in orange. Bottom: log-odds ratios obtained from the contingency table comparing cluster assignment and DTC labelling in vitro (left) or in vivo (right) across cells at T0. **e** Association between parental state (T0) and clone expansion in vivo, without and with treatment. Top: bar plot showing the relative normalised abundance of T0 clusters in every untreated (left, as reported in Fig. 2f) or treated tumour (right), respectively (unassigned clones shown in grey). Bottom: cartoon highlighting the subpopulations enriched in DTCs (odds ratio values reported below) [**a** Created with Biorender.com released under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license].

https://doi.org/10.1038/s41467-024-51424-4

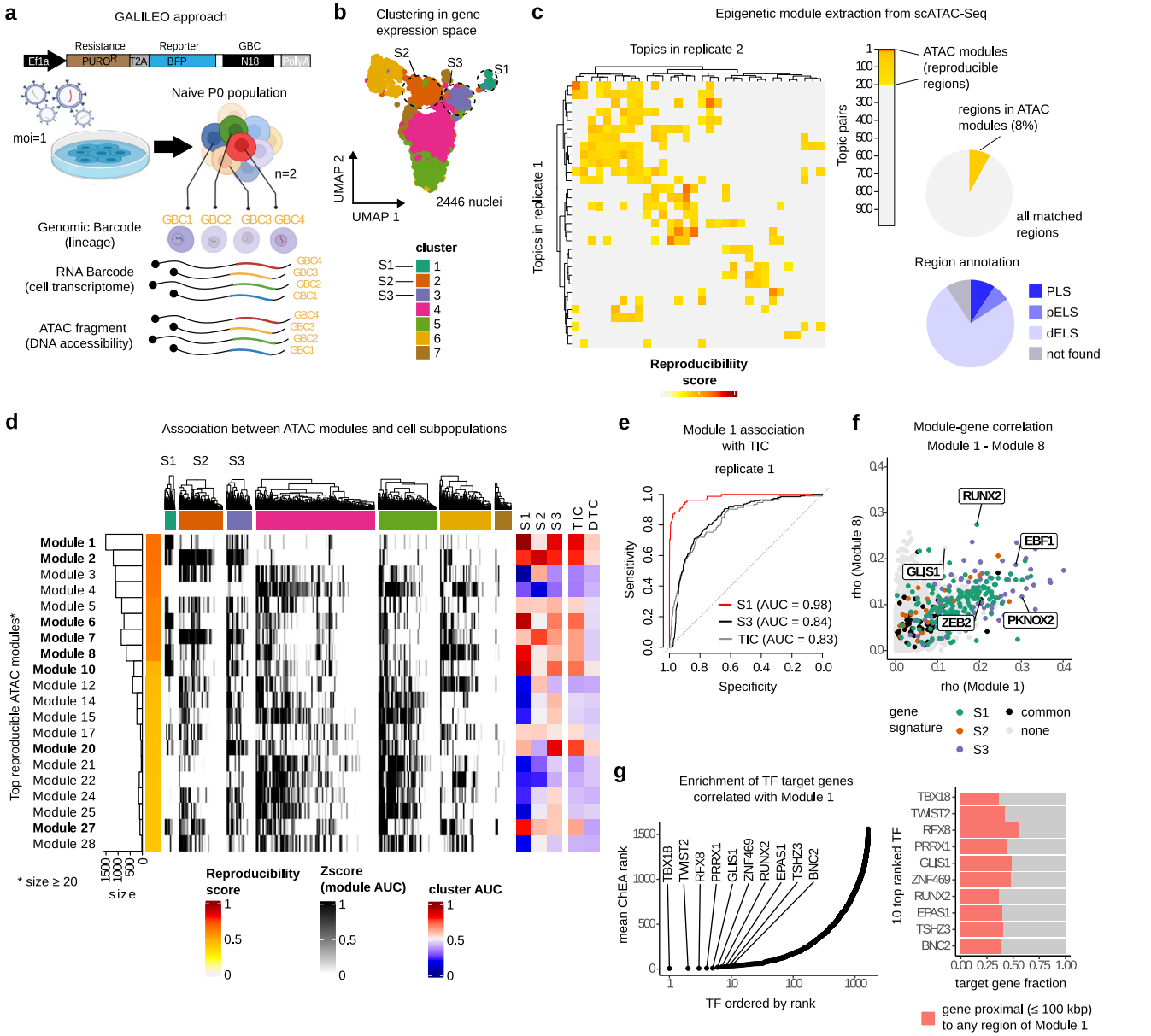


Fig. 4 | A major DNA accessibility programme predicts tumour initiation and is associated with S1 and S3 subpopulations. a Genomic bArcoding pLus sIngLE-cell multi-Omics (GALILEO) strategy. Barcoded SUM159PT cells are processed with Multiome. The readout of each cell is lineage, gene expression profile and DNA accessibility state. **b** UMAP representation in gene expression space coloured by cluster (replicate 1, 2446 nuclei). **c** Topic modelling on ATAC-seq regions. Left: comparison of the output of topic modelling on the two replicates; and entries are reproducibility scores (see section "Methods"); topics highly correlated with coverage are not shown (see Supplementary Fig. 10b). Centre: topic pairs ordered by non-increasing score (yellow: non-empty modules). Right: reproducible region fraction and their annotation by ENCODE registry (bottom) of candidate *cis*-regulatory elements (cCREs) as PLS (promoter), pELS (proximal enhancer), or dELS (distal enhancer) (replicate 1). **d** Comparison between ATAC modules and gene expression in single nuclei (replicate 1). Rows are the top 20 scoring modules, columns are nuclei and entries are module AUC scores representing the overall accessibility of a module. The association (AUC) of a module to subpopulations (S1, S2, and S3) and cancer fates (TIC and DTC) in vitro is shown on the right; high associations (AUC > 0.75) with any subpopulation are highlighted in bold. Columns are clustered on Euclidean distances using a complete method from hclust package. **e** Module 1 AUC as a predictor of S1 (red), S3 (black), or TICs (grey) on replicate 1. **f** Association between module AUC scores and gene expression (replicate 1). Each dot is a gene and its value represents the (positive) Spearman's ρ correlation coefficient between its expression and module AUC score. Genes are coloured according to scRNA-seq and Multiome gene signatures. Transcription factors with $\rho \geq 0.2$ in either module are labelled. **g** TF enrichment on the genes whose expression is positively correlated with Module 1 AUC. Left: TFs sorted by rank, with top 10 ranked TFs labelled. Right: fraction of genes (coloured in pink) whose locus is ≤100 kbp away from any region in Module 1, for the 10 top-ranked TFs [**a** Created with Biorender.com released under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license].

(AUC = 0.84) (Fig. 4e). Module 20 (50 regions) showed an equally strong association with S3 (AUC = 0.84). To further assess the relationship between ATAC modules and subpopulations, we correlated genome accessibility (using module AUC, see above) with gene expression across cells (Spearman's ρ; Fig. 4f and Supplementary Data 12). This approach ensures that mechanisms involving either *cis* or *trans* genomic elements can be captured, as no constraint on region-gene proximity was used. Overall, module AUC was positively correlated with the expression of genes of the associated subpopulation signature (Fig. 4f and Supplementary Fig. 10d). These results support the hypothesis that ATAC modules contain regulatory elements jointly involved in the control of specific transcriptional programmes.

https://doi.org/10.1038/s41467-024-51424-4

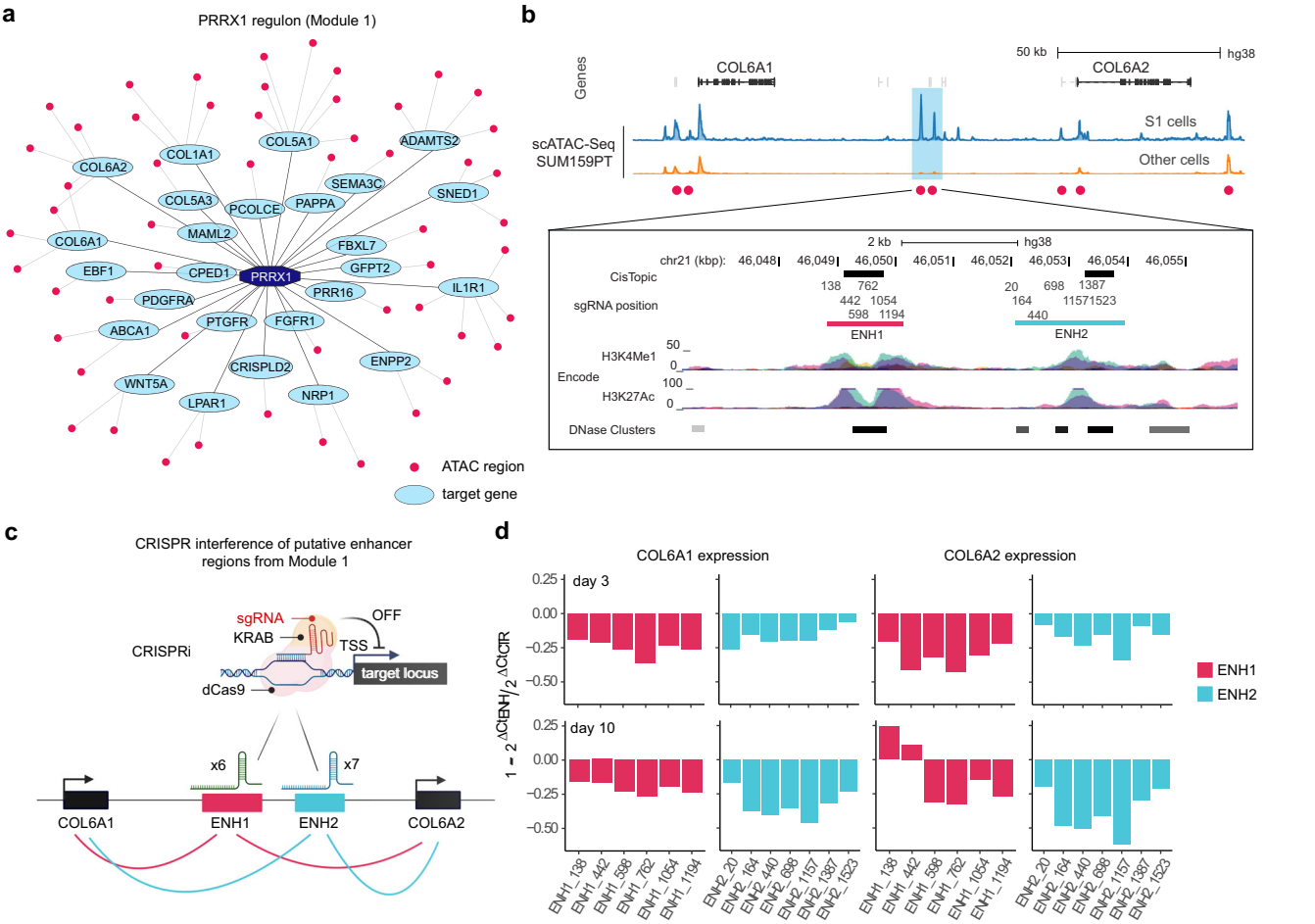


Fig. 5 | Silencing of selected Module 1 regions triggers a reduction in transcription of proximal genes. a The "PRRX1 regulon" includes the set of genes whose expression is positively correlated with Module 1 AUC and that (i) are predicted as PRRX1 targets by ChEA3 and (ii) lie at ≤100 kbp from any region in Module 1. **b** *COL6A1* and *COL6A2* loci; shown are the scATAC-seq peaks (aggregate signal, replicate 1). Red dots label Module 1-specific regions. In the magnification is shown the region containing the two enhancers, ENH1 and ENH2, together with ENCODE regulatory tracks for H3K4Me1, H3K27Ac, and DNase clusters and the position of sgRNAs for CRISPRi (see also **d**). **c** Scheme showing the CRISPRi approach and the model of Enhancer-Gene pair regulation. **d** RT-qPCR data showing the impact in *COL6A1* and *COL6A2* expression in TM4SF1^{high} SUM159PT_KRAB cells observed upon expressing sgRNAs targeting ENH1 (in pink) or ENH2 (in blue) (see also **b**). Data are measured at day 3 (top) or day 10 (bottom) after dCas9-KRAB induction by doxycycline ($n = 6$ and 7 sgRNAs, respectively) and compared to a group of non-targeting sgRNAs ($n = 4$) [**c** Created with Biorender.com released under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license].

Tumour-initiating clones share a common chromatin priming state

We then sought to investigate the role of ATAC modules in cancer phenotypes, namely, tumour-initiation and drug-tolerance capacity. Importantly, Module 1, the top reproducible accessibility state, predicted TICs with high specificity and sensitivity (Fig. 4e and Supplementary Fig. 10e), consistently with it being associated with S1 and S3, which, in turn, are enriched in TICs (see also Fig. 2e, f). This suggests that the tumour-initiating capacity may be linked to a specific and pre-existing epigenetic state and may explain the phenotypic relationship between the cells of S1 and S3. Subsequently, we used transcriptional and epigenetic information jointly and at single-cell resolution to highlight the gene regulatory networks and the epigenetic determinants involved in the tumour-initiating capacity. For each module m, we used the set G_m of positively correlated genes (see above) as input to measure the transcription factor activity in m given a set of putative targets for each TF[57] (Fig. 4g and Supplementary Data 13). Among the top-ranked TFs for Module 1, we detected several TFs that have been previously linked to tumour-initiation capacity, including TWIST2, PRRX1, and RUNX2; note that the S1 gene signature is enriched in RUNX2 targets (see Supplementary Fig. 4h). TWIST2 is a member of the TWIST family of TFs, which has been extensively associated with poor tumour prognosis, EMT, and stem-cell activity in breast cancer[58–60]. Similarly, RUNX2 activity has also been linked to the regulation of EMT, matrix remodelling, and invasive phenotypes[52], which lead to metastasis, notably in breast cancer[35,61]; finally, PRRX1 has been recently shown to sustain metastatic dissemination and induce a switch to a mesenchymal-like state in a melanoma cancer model[26]. RUNX2, TWIST2, and PRRX1 were top ranked in Module 1 also using the set of genes proximal to Module 1 ATAC regions as input (Supplementary Fig. 10f). For each TF, we identified its regulon by the set of positively correlated target genes whose locus is proximal (<100 kbp) to any region of Module 1. Several genes in the S1 signature, including procollagen C-endopeptidase enhancer 1 (*PCOLCE*) and collagen-encoding genes (*COL6A1*, *COL6A2*, *COL5A1*, *COL5A3*), were found as part of PRRX1 and TWIST1 regulons (Fig. 5a and Supplementary Fig. 10g). To directly verify the role of the genomic elements of Module 1 in gene regulation, we selected two regions located in the proximity of COL6A1 and COL6A2, highly accessible in S1 cells (Supplementary Fig. 10h) and classified as dELS by ENCODE cCRE (Fig. 5b, top). Subsequently, we targeted the two regions by means of an inducible CRISPR interference strategy[62,63] (Fig. 5c). We observed that repression of either region led to a consistent and reproducible reduction in *COL6A1* (up to 36% and 46%) and *COL6A2* expression

(up to 43% and 62%; Fig. 5d), both at early (3 days) and late time points (10 days) post-dCas9-KRAB induction.

A subset of the drug-tolerant subpopulation exhibits a pre-existing genomic amplification

Module 20 predicts S3 with high specificity and sensitivity (AUC = 0.84 and 0.75 in the two replicates; Fig. 6a and Supplementary Fig. 11a). As shown in Fig. 3, the S3 programme is associated with increased drug tolerance, both in vitro and in vivo. We noticed that most regions of Module 20, as well as several genes of the S3 signature, were located on a 5.5 Mbp-long region of chromosome 11 (Supplementary Fig. 10i). The odds are low that an epigenetic regulatory mechanism involves a cluster of highly localised genomic elements and thus we reasoned that a genetic alteration might better explain the transcriptional profile of S3. When interrogating the whole-exome sequencing profile of paclitaxel-treated samples and comparing it to that of untreated cells ($n = 3$; see Fig. 3a), we detected 18 recurrent copy-number variants (CNVs; Fig. 6b and Supplementary Fig. 11b). Notably, the top amplified region (average $\log_2(FC) = 0.53$ and p value = 5.65e − 185) lied on chromosome 11, specifically across bands 11q23–11q24 (Fig. 6c). These results suggest that the amplification was already present in S3 cells before treatment and that Module 20 captures a specific genetic background of S3, rather than a localised increase in chromatin accessibility. This implies that the drug tolerance phenotype is, at least in part, genetically determined, and, in turn, suggests that a subset of DTCs could maintain a stable memory of the treatment. Therefore, we investigated the susceptibility of cells to paclitaxel upon a recovery period of 24 days after a first round of treatment (see section "Methods", Fig. 6d, top, and Supplementary Fig. 11c). Upon a second round of treatment, clonality was reduced by 63% (average recovery at day ≥ 17 compared to day ≤ 3; Fig. 6d, bottom), suggesting that chemotherapy was still effective; however, drug sensitivity was low compared to a single round of treatment, where only 19% of clones survived. Finally, to verify the specificity of the association observed between the chr11 region amplification and resistance to paclitaxel, we examined a drug resistance model, where SUM159PT cells were repeatedly treated with increasing doses of paclitaxel (see section "Methods", Fig. 6e, top, and Supplementary Fig. 11c) up to the onset of a drug resistance phenotype. The WES profile confirmed an amplification on chromosome 11 whose locus overlaps 75% of the above-detected one (Fig. 6e, bottom, and Supplementary Fig. 11d and Supplementary Data 14). We conclude that, in SUM159PT, paclitaxel-based chemotherapy causes a clonal expansion of a subpopulation harbouring the amplification of 11q23–11q24.

Two distinct clone lineages can enter a drug tolerance state

We note that the 11q23–11q24 amplification is only found in S3. However, in vitro, many DTCs fall outside S3 (see Fig. 6a and Supplementary Fig. 8a), highlighting their molecular heterogeneity at baseline. To further characterise the pathways leading to drug tolerance in vitro, we designed a single-cell time-course experiment on barcoded SUM159PT cells (Fig. 7a, top). To minimise technical variability, we treated cells for 3 days with paclitaxel, collected samples every 2 days after drug removal (days 5–15) and sequenced them simultaneously using a reverse time course as in ref. 64. Consistent with untreated samples, we assigned a clonal identity to most of the high-quality cells (80–84%, Supplementary Fig. 12a, b). Treatment elicited a progressive clone selection, from 1126 distinct clones at day 5 to 433 at day 15 (Fig. 7a and Supplementary Data 15). An independent experiment confirmed the clone selection dynamics (Supplementary Fig. 12b, c). Chemotherapy induced a substantial transcriptional change (Fig. 7b, c, and Supplementary Fig. 12c, d and Supplementary Data 16). At initial time points (d5–d9), most cells were drug-sensitive and developed two distinct responses to treatment: the one characterised by the induction of stress response pathways, including amino-acid deprivation, unfolded protein response, and inflammation (cluster 1); the other mostly characterised by autophagy (cluster 2; Fig. 7d, and Supplementary Fig. 12e and Supplementary Data 17). At late time points, surviving cells showed enhanced translation activity (cluster 3). At initial time points (d5 and d7), cells belonging to the surviving pool (i.e., DTCs) accounted for only 9% of the whole sample and their transcriptional profile was scattered (Fig. 7b and Supplementary Fig. 12d). DTCs survived the treatment by remaining in a state of suspended proliferation for several days and, starting from days 9 to 11, entered an intense proliferation phase; from day 13 on, DTCs constituted 73%–85% of surviving cells (Fig. 7a and Supplementary Fig. 12c) and acquired a distinguishing transcriptional profile (Fig. 7e). At late time points (days 13–15), cells stemming from the same clone at baseline overall displayed a divergent transcriptional programme (solid line), comparable to that of cells belonging to different clones (dashed line, Fig. 7f). Then, we asked whether any distinguishing transcriptional footprint exists in highly expanded clones. To do this, we devise an unsupervised approach to find sets of mutually similar clones (by defining a *pair propensity* score across gene expression neighbourhoods; see section "Methods" and Fig. 7g). Two groups of clones, or *lineages*, showed mutually high transcriptional similarity (Fig. 7g, h and Supplementary Fig. 13a), suggesting that multiple pathways to drug tolerance may exist in SUM159PT cells. Both lineages contained highly abundant clones, indicating that the transcriptional readout is not associated with proliferation potential. The two lineages were reproducible both across time points and across independent experiments (Supplementary Fig. 13b, c). On average, they accounted for 50% (lineage 1) and 35% (lineage 2) of the cells at late time points (the remainder fraction belongs to unclassified clones). Notably, clones in lineage 1 stemmed from S3 and were characterised by a pre-existing genomic amplification on band 11q23–11q24. Among the upregulated genes in lineage 1, we detected *MT1E*, whose relevance in breast and other cancer types has been extensively proven[65,66], as well as *FEZ1* and *RPS25*, top significant genes in the S3 signature (Fig. 7i, j and Supplementary Data 16). Consistently, genes located within the 11q23–11q24 amplification, which is specific to S3, were upregulated in clones belonging to lineage 1 (Fig. 7k and Supplementary Fig. 13d, e). This showed that the transcriptional differentiation breadth of the two lineages is determined by the genetic background of the ancestor clone, specifically, depending on whether it carries the 11q23–11q24 amplification or not. In contrast, we did not detect any lineage 2-specific copy-number aberration (Supplementary Fig. 13f). The top upregulated genes in the lineage 2 signature, namely, *S100A2*, *IGFBP2*, *IFI27*, and *PVT1*, were most highly expressed immediately following treatment in both DTCs and non-DTCs (Fig. 7j), suggesting that the transcriptional programme of lineage 2 is not DTC-specific. The early response to paclitaxel includes upregulation of *PVT1*, a long non-coding gene acting as a negative regulator of the transcription factor MYC, a key regulator of growth and cellular metabolism, frequently associated to breast cancer[67,68]. Consistently, we observed that MYC activity decreased immediately after treatment and increased during adaptation in response to it (Supplementary Fig. 13g), with only slight differences between the two lineages (Supplementary Fig. 13h). Consistent with our findings, recent evidence showed that reduced MYC activity promotes a chemotherapy survival phenotype in breast cancer via the adoption of an embryonic-like diapause state[69]. In conclusion, we mapped the transcriptional response upon drug treatment at a clonal level and isolated different pathways of cancer transcriptional evolution leading to resistance, one of them being invariably linked to a pre-existing genetic rearrangement.

Discussion

One of the main challenges in cancer biology is predicting how tumours evolve in response to changes in the tumour environment. The capacity of one or more clones to sustain tumour growth at distal sites or to trigger disease relapse upon cytotoxic treatment may

https://doi.org/10.1038/s41467-024-51424-4

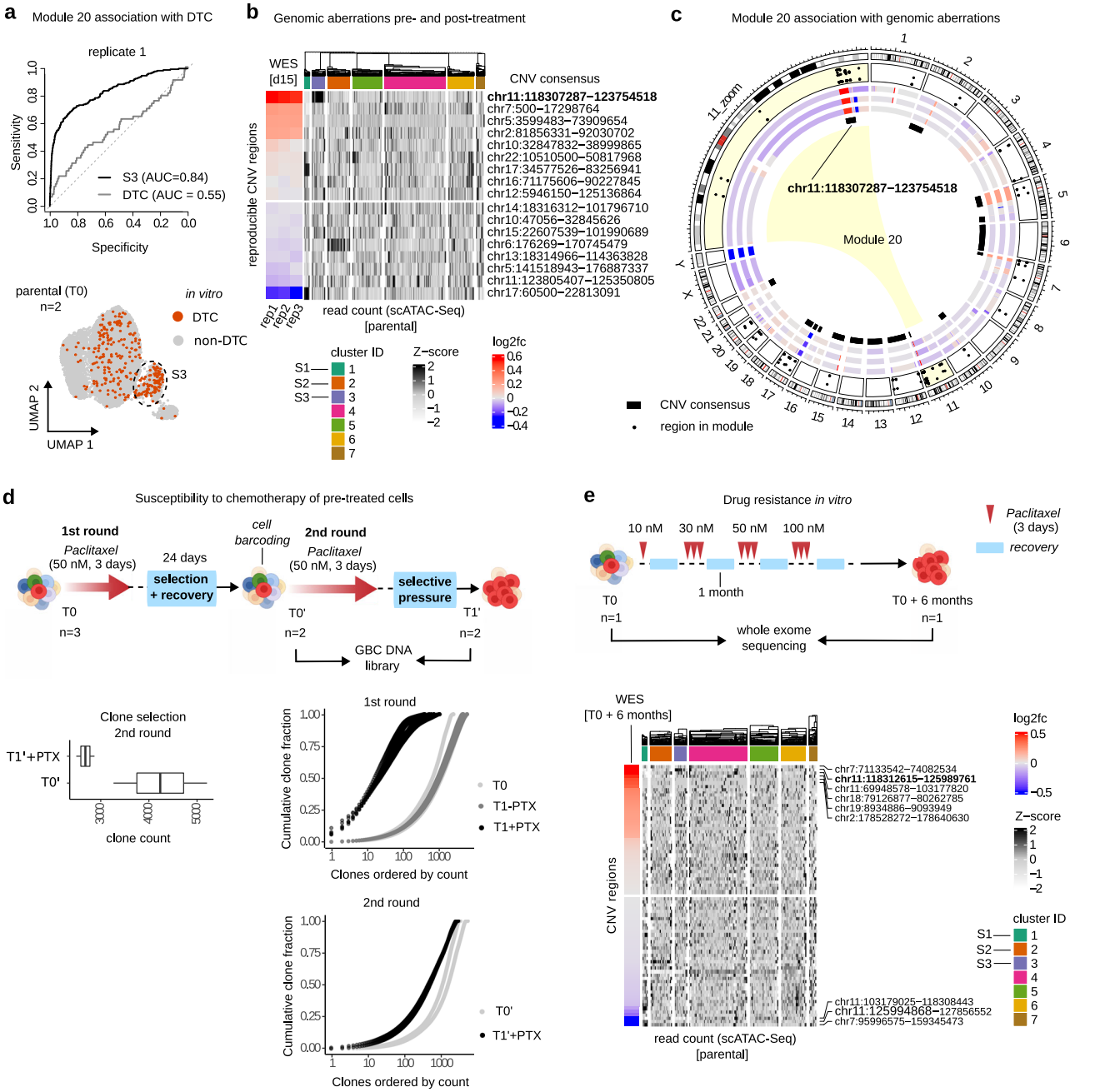


Fig. 6 | Drug-tolerant clones display chr11 amplification. a Module 20 AUC as a predictor of S3 (black) or DTCs in vitro (grey) on Multiome replicate 1. Bottom: UMAP representation of T0 cells on gene expression space coloured as in Fig. 3d. **b** Copy-number variants (CNVs) pre- and post-treatment (drug tolerance assay, replicate 1). ATAC-seq coverage (Multiome) on CNV loci[52]. Rows are consensus CNV from paclitaxel-treated samples (day 15; $n = 3$; experiment as in Fig. 3a); columns are nuclei in Multiome replicate 1; entries are cumulative ATAC counts per CNV locus per nucleus. The coverage $\log_2$(FC) between each treated sample and the parental (left) and chromosome location (right) are indicated. Columns are clustered with a complete method on Euclidean distances using hclust. **c** Circos plot showing the association between Module 20 regions at baseline (see Fig. 4c) and CNVs in treated condition. In the dot plot, the *y* axis is the IDR for the regions in Module 20; CNVs for each replicate are coloured by $\log_2$(FC); consensus CNVs are shown in black. **d** Drug tolerance assay, second round. Top: SUM159PT cells were treated with paclitaxel and clone selection was stabilised until T0′; cells were subsequently infected with the GBC library sorted by BFP expression, subjected to a second round of treatment, and harvested at T1′. The three populations (T0, T0′, and T1′) were sequenced. Bottom left: total clone count at T0′ and T1′ (bounds of box: upper (q75) and lower (q25) quartiles; centre: median; upper whisker: min{max(*x*), q75 + 1.5·IQR}; lower whisker: max{min(*x*), q25 − 1.5·IQR}). Bottom right: cumulative clone distribution after either one round (see also Fig. 3b, top) or two rounds of treatment. **e** Top: long-term drug resistance assay in vitro. SUM159PT cells were repeatedly treated with increasing doses of paclitaxel until resistant clones were obtained, which were then processed by WES. Bottom: ATAC-seq coverage (Multiome) on CNV loci[52], as in (**b**) [**d**, **e** Created with Biorender.com released under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license].

depend on a specific set of molecular characteristics. Their identification has been the focus of intense research efforts, both for the clinical applications and for the understanding of the mechanisms underlying tumour plasticity.

Recently, it has been suggested that the cancer stem-like pool might be heterogeneous, with distinct subpopulations primed to different fates[4–10]. The SUM159PT model is one representative example, in which distinct states coexist in equilibrium[33]. Here, we provided the

https://doi.org/10.1038/s41467-024-51424-4

distinctive transcriptional and epigenetic traits of each sub-pool. We analysed tumour initiation and drug tolerance at single-cell level on lineage-barcoded cells and on the same system, providing a high-resolution representation of tumour complexity.

We initially asked which clonal subpopulations lie in a defined transcriptional or epigenetic state. In the case of SUM159PT, only a fraction of clones displayed a stable transcriptional profile at the subpopulation level, hinting that these may encode for specific functions. Indeed, the TICs were almost exclusively associated with either S1 or S3 signatures, with S1 showing the strongest enrichment and S3 also encoding for DTCs. The third signature (S2) was not attributable to any cancer property, although it contains basal markers (e.g., *miR-205HG*, *HMGA1*).

An important and direct conclusion of our study is that TICs and DTCs do not coincide but coexist in the same cancer population, sharing a minor subset of clones (belonging to S3). TICs and DTCs can significantly change their transcriptional profile during cancer evolution, adapting to the environment and conditions. In transplanted tumours, TICs lose their baseline signature upon expansion and maintain upregulation of only a handful of markers (e.g., *S100A4*, *TM4SF1*), whose phenotypic relevance is confirmed in the literature[20,37,39,52]. Similarly, DTCs undergo a massive transcriptional

Fig. 7 | Drug-tolerant clones stem from two distinct lineages. a Time-course drug tolerance assay (exp1). Top: SUM159PT cells were treated with paclitaxel for 3 days, harvested every 2 days post-paclitaxel removal ($n = 1$/time point), and processed with scRNA-seq using reverse time-course multiplexing. Bottom: fraction of DTC in vitro (see Fig. 3c, top) across time points. **b** Drug-tolerant clone selection. In total, 7884 treated cells are shown and coloured by clone class. **c** UMAP representation of cells at days 5, 7, 9, 11, 13, and 15 and coloured by cluster. **d** Cluster signatures: rows are genes, columns are clusters, and entries are $\log_2$(FC) values between a cluster and its complement. Significantly upregulated genes in any cluster are shown and order by average log2(FC) and adjusted p value ranking (lower to higher; MAST method, Bonferroni correction). The top 10 genes for each cluster are labelled. **e** Gaussian kernel density of Euclidean distances in gene expression between DTCs, non-DTC, and between DTCs and non-DTC, before (T1, left) or after treatment (exp1, right). **f** Gaussian kernel density of Euclidean distances between sister and non-sister cells at day 15. **g** Gene expression similarity bias by pair propensity. Left: calculation of pair propensity between clones (see section "Methods"). Right: pair propensity for top expanded clone pairs at day 15 (clones i with $p_{ii} < 1$ not shown); clones are annotated with clone abundance (count per thousand cells, rows) and lineage (columns). **h** UMAP representation of cells at days 13 and 15 coloured by lineage assignment (unassigned clones in grey). **i** Rows are lineage gene signatures, columns are independent experiments, and entries are average $\log_2$(FC) values between lineages 1 and 2 after treatment. Genes are ordered by non-increasing $\log_2$(FC) in exp1. The 10 DEGs with higher and lower $\log_2$(FC) are labelled. **j** UMAP plot of cells at day ≥ 5 coloured by log-normalised gene expression of the four top $\log_2$(FC) genes of lineages 1 and 2. **k** Rows are genes whose locus lies in chr11:118307287–123754518 (see Fig. 6b), columns are cells at day 15 and entries are scaled log-normalised UMI counts. Rows are sorted by non-increasing average expression; columns are clustered with complete method on Euclidean distances.

reprogramming after treatment and, thus, are strikingly distinct from their non-DTC counterpart, a behaviour observed in other cancer cell models[64,69]. Using lineage tracing information alongside a time-course single-cell profiling, we could describe two distinct and co-occurring transcriptional trajectories in drug adaptation. Differently from TICs, the DTC subpopulation did not show a strong transcriptional or epigenetic determinant at the baseline. However, a subset of DTCs, lying in S3, shows amplification of a 5.5 Mbp-long region of chromosome 11 (bands 11q23–11q24). This genetic background also reproducibly segregated with chronic treatment resistance in the SUM159PT model. To our knowledge, this amplification is not reported as a recurrent alteration in cancer (https://cancer.sanger.ac.uk/cosmic). We confirmed the existence of a large amplicon spanning the chr11:118M–126M region in a small fraction of primary breast tumour samples from the TCGA dataset (73 out of 1084 cases; cBioPortal), but we could not assess its association with resistance to chemotherapy treatments, nor the role in the phenotype of the individual genes lying in the locus. However, among the genes lying in the chr11:118M–126M region is *MIR100HG*, a long non-coding RNA that encodes for miR-100 and miR-125b, the latter being a miRNA known to confer resistance to taxol treatment in TNBC cell lines[70,71].

A key innovative element of our study is the use of a cutting-edge approach combining single-cell multi-omic profiling (transcriptome and DNA accessibility) with lineage tracing (clone information at an arbitrary time t, which we call the baseline—P0). Specifically, we found a putative regulatory programme (Module M1) common to both S1 and S3, which elegantly links the two tumour-initiating states. Moreover, the inferred transcription factors and the corresponding regulons comprise both genes and regulatory regions, including non-coding components (lncRNAs and enhancers). Note that these elements might be detectable in bulk experiments, but only the single-cell resolution can explain their relationship, which may be subpopulation-dependent. The TF hubs of the predicted regulons are fully supported by the literature; for instance, PRRX1, TWIST2 and RUNX2 have been linked to breast cancer and to the EMT, a founding element shared by both tumour aggressiveness and stem cell identity programmes[52,72,73].

An important question is whether the molecular traits distinctive of TIC are specific to SUM159PT or are generalisable. In large breast cancer datasets, the upregulation of S1 and S3 signature genes predicts basal features, which are typically associated with cancer stemness[74]. S1 signature genes have also been found as upregulated within cell subpopulations in other breast cancer models (MDA-MB-231 TGL), as well as associated with general cancer meta-programmes across over 1000 primary tumour samples[51]. Of note, the strongest association was found with the hybrid EMT meta-programme, which shares several markers with S1, including TM4SF1, used to enrich for functional TICs in SUM159PT (in this work) and in other experimental models, such as MDA-MB-231, murine 4T1 breast cancer cells, and MMTV-Neu tumours[53,54]. Furthermore, by employing either lineage tracing or single-cell omics, recent literature highlighted programmes with remarkable similarities to the ones we reported. In a mouse model of metastatic pancreatic cancer, Simeonov et al. highlighted a hybrid EMT state in metastatic dissemination which shares S100 family gene expression (see Supplementary Fig. 4g) and is predictive of reduced survival in both pancreatic and lung cancer patients[20]. In a genetically engineered mouse lung cancer model, a co-accessibility module characterised by RUNX2 activity was identified and linked with the acquisition of a pre-metastatic state and the outcome of human lung cancer[52]. Interestingly, the S1 signature genes that are also significantly associated with the hybrid EMT meta-programme are putative RUNX2 targets (see Supplementary Fig. 4h). In line with this observation, the KP-tracer approach for in vivo lung cancer lineage tracing allowed to show that tumours evolve through stereotypical trajectories, with the transient activation of cellular plasticity programmes and a subsequent clonal sweep of highly fit subpopulations marked by an early or late mesenchymal transition[27]. Finally, single-cell lineage tracing revealed the underlying programme of a pool of metastatic initiating cells in melanoma, characterised by high PRRX1 expression and promoting the establishment of a mesenchymal-like cell state[26].

Our and published evidence suggest a model where the mechanisms influencing clonal fate and tumour evolution tend to converge towards a common epigenetic state, often established before a challenge and, therefore, predictable. The main programme we identified was a hybrid EMT and the associated transcription factors (e.g., RUNX2 and TWIST). It is worth noting that one of the results obtained with GALILEO approach is the precise reconstruction of the network of genes and specific regulatory elements related to the above programme in TNBC cells. Indeed, these elements can potentially highlight cancer dependencies with clinical impact. We foresee that combining cutting-edge molecular tools at the genome scale, like the ones presented here, as well as genetic ones, with suitable computational frameworks, could critically contribute to further dissect the role played by different transcriptional, epigenetic and genetic layers in cancer evolution. Our study has made it clear that a multi-layered framework is feasible and an invaluable resource to this end. Finally, our work directly shows that both genetic and epigenetic mechanisms can promote cancer evolution towards specific fates, and that these mechanisms may coexist in the same tumour within specific cell subpopulations.

Methods

Mice

All animal studies were conducted with the approval of the Italian Ministry of Health and in compliance with the Italian law (D.lgs. 26/2014), which enforces Dir. 2010/63/EU (Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes) and EU 86/609 directive. Proper permit and consent were granted (Protocol No. 779/2020) by the institutional organism for ethics and animal welfare on

experimental procedures (OPBA, Cogentech). Animals were bred and maintained under pathogen-free conditions in a controlled environment (18–23 °C, 40–60% humidity and with 12-h dark/12-h light cycles) at the certified Cogentech mouse facility located at IEO/IFOM campus (The FIRC Institute of Molecular Oncology, Milan, Italy).

In vivo xenograft

Immunodeficient NOD.Cg-PrkdcscidIL2rgtm1Wjl/SzJ (also known as NOD/SCID/IL2R$\gamma_c^{-/-}$) mice were anesthetised by intraperitoneal injections of 1.25% solution of tribromoethanol (0.02 ml/g of body weight). Barcode-bearing SUM159PT cells were resuspended in a mix of 14 µl PBS and 6 µl Matrigel and implanted in the fourth inguinal mammary gland of 10-week-old animals. Mice were monitored twice a week by an investigator. For the chemotherapy treatment studies, tumours were allowed to grow to palpable lesions (~20–30 mm^3), then mice were randomised into groups and each group was treated intraperitoneally with either paclitaxel (PTX) (10 mg/kg in PBS) or vehicle (PBS) every 5 days for a total of three injections. Mice were euthanised according to our experimental protocol and institutional guidelines when tumours euqalled 1.2 cm in their largest diameter. Maximal tumour burden was not exceeded. Tumour growth dynamics were monitored every 3 days by calipers measurements. For in vivo limiting dilution assay transplantation experiments, decreasing dilutions (1:10,000; 1:1000; 1:100) of SUM159PT were resuspended in a mix of 14 µl PBS and 6 µl Matrigel and transplanted in the fourth inguinal mammary gland of 10-week-old animals. Animals were monitored as before and euthanized after 1.5–3 months (depending on tumour latency).

Tissue harvest and processing

The primary tumours were removed when the tumour reached an approximate diameter of 1.2 cm. The animals were anesthetised with tribromoethanol, and the tumours were resected. The solid tissue was rinsed with PBS, minced with scalpels, followed by mechanical dissociation using gentleMACS (Miltenyi) in a digestion mix (collagenase, hyaluronidase, 5 µg ml^{-1} insulin, Hepes, hydrocortisone). The cell suspension was incubated for 25′ at 37 °C followed by an additional step of gentleMACS dissociation. Following a wash with base medium the cells were consecutively passed through 100, 70 and 40 µm filters. Primary tumour cells were treated with ACK lysis buffer (Lonza) followed by resuspension in 1% BSA/PBS and processed using the Mouse and Dead Cell depletion kits according to the manufacturer's directions (Miltenyi).

Cell cultures

We maintained HEK293T and MDA-MB-231 TGL cells in Dulbecco's Modified Eagle's Medium (DMEM) with 10% of TET-FREE foetal bovine serum (FBS) and 1% penicillin–streptomycin. Cells were grown in a humidified atmosphere at 5% CO_2 at 37 °C. SUM159PT cell line and derivatives (Asterand) were cultured in Ham's F12 medium with 5% TET-FREE FBS, human insulin (5 µg/ml), hydrocortisone (1 µg/ml), and Hepes (10 mM). Barcoded SUM159PT and CRISPRi cell line medium was supplemented with 2 µg/ml puromycin for selection. Cells were grown in a humidified atmosphere at 10% CO_2 at 37 °C.

Perturb_SUM159PT cell line generation

Perturb-seq GBC library[18,75] was a gift from Jonathan Weissman (Addgene ID #85968). The library contains a random 18-nt guide barcode (GBC) close to the polyadenylation signal of the blue fluorescent protein (BFP). The estimated complexity of the library is >5 million unique GBCs. We amplified the Perturb-Seq GBC library in *Escherichia coli* (*E. coli*) ElectroMAXTM DH5α-ETM electro-competent cells (Thermo Fisher Scientific), as indicated by the authors[75]. DNA extraction was performed with NucleoBond Xtra Maxi (Macherey-Nagel) kit according to the manufacturer's instructions. Viruses were produced in HEK293T at 80% of confluency with the following transfection mix: 20 µg Transfer Vector, 13 µg of psPAX2 (gag&pol) (Addgene #12260), 7 µg of pMD2G (envelope) (Addgene #12259), 94 µL of RT $CaCl_2$ and water up to 750 µL. Then, 750 µL of 2xHBS were added dropwise and 500 µL/10 cm plate of transfection mix was added to the cells. The medium was changed after 6 h and viruses harvested after 48 h, filtered (0.22 µm) and ultracentrifuged for 2 h at 50,000 × *g* (rotor SW32 Ti, Beckman Coulter), at 4 °C. The viral pellet was resuspended in 300 µL of PBS 1X and stored at −80 °C. To produce the Perturb SUM159PT cell line for lineage tracing experiments, 75,000 cells were seeded and infected with an estimated MOI of 0.1 in the presence of 8 µg/ml of polybrene (Sigma-Aldrich). After selection, transduction efficiency was measured by FACS analysis, which revealed that 8.6% of cells were successfully infected.

SUM159PT_KRAB cell line generation and CRISPRi experiments

For the generation of the PB-TRE-dCas9-KRAB plasmid, the DNA sequence of KRAB repressor domain was amplified by PCR from the pHAGE TRE dCas9-KRAB (Addgene plasmid #50917) and cloned in frame into the PB-TRE-dCas9-VPR backbone (Addgene plasmid #63800) within the AscI/AgeI sites. The cloning was sequence-verified by Sanger sequencing. SUM159PT cells were transfected in MW6 plates, following Lipo3000 transfection protocol (ThermoFisher Scientific) with 500 ng of transposon DNA (PB-TRE-dCas9-KRAB) and 200 ng of SuperPiggyBac transposase helper plasmid (SystemsBioscience). After at least 72 h from transfection, cells were selected with 200 µg/ml Hygromycin B. The PB-TRE-dCas9-KRAB SUM159PT cell line is referred in the text as SUM159PT_KRAB. We expressed sgRNAs upon cloning into lentiGuide-Puro sgRNA backbone (Addgene #52963) within BsmBI (Esp3I) sites. Lentiviruses were generated in six-well plates, following the Lipofectamine 3000 (Thermo Fisher) protocol. The transfection was performed by mixing the construct of interest, psPAX2 (gag&pol)(Addgene #12260) and pCMV-VSV-G (envelope) (Addgene #8454) plasmids at a ratio of 4:3:1. Viruses were collected after 24 h, filtered and frozen. We generated stable cell lines expressing single sgRNAs by lentiviral infection of 150,000/well SUM159PT_KRAB cells. Lentiviral supernatants (1:3 dilution) were added to cells, supplemented with 1 µg/ml of polybrene. After 24 h, cells were selected with 2 µg/ml puromycin (Gibco). For the CRISPRi experiment, we plated sgRNA-expressing stable cell lines with 100 ng/ml of doxycycline and harvested cells after 3 days.

Paclitaxel treatment

A stock aliquot of paclitaxel (obtained from the IEO hospital) was prepared (70 µM) in PBS and used to treat cells. The dose of paclitaxel used for in vitro experiments was established by a dose–response curve of SUM159PT treated for 72 h. IC50 concentrations were estimated by parallel fit estimation (JMP software, $n = 4$). For short-term, single-treatment experiments, paclitaxel was used at the final concentration of 50 nM (~IC95). After 3 days of treatment, the medium was changed every 3 days without adding the drug. For the analysis of the susceptibility of pre-treated cells (shown in Fig. 6d), SUM159PT cells were treated with paclitaxel (50 nM) and surviving clones were allowed to recover for 21 days. Subsequently, cells were infected with the GBC library, sorted by BFP expression and subjected to a second round of treatment (50 nM). For the generation of the drug resistant model (shown in Fig. 6e), SUM159PT cells were treated multiple times with increasing doses of paclitaxel (10, 20, 50, and 100 nM). Drug-adapted cells were able to survive even when treated with 100 nM paclitaxel and were collected 90 days after treatment for WES analysis.

TM4SF1^{high} flow cytometry

SUM159PT_KRAB cells were stained with anti-TM4SF1-APC (clone: REA851 Miltenyi Biotec) antibody for 10′ at 4°C, in the dark and sorted using Fusion Aria Sorter. The bulk population was FAC-sorted using FSC/SSC gate, while for the TM4SF1^{high} subpopulation different gating

strategies were tested (top 5% or top 10% APC fluorescence intensity as shown in Supplementary Fig. 5d); top 5% showed the best enrichment for S1 markers (shown in Supplementary Fig. 6b) and was used in subsequent experiments. After passive propagation in vitro, 9 days after sorting, both populations were transplanted into NOD/SCID/IL2Rγc, mice and tumours were collected as described above.

RNA sequencing. RNA was extracted from mouse-depleted and dead-cell-depleted samples. The bulk and TM4SF1^{high} sorted cells were either immediately processed by RNA-seq or passively propagated in vitro for 43 days (corresponding to passage 19). Cells after 9 and 43 days from sorting were also collected and finally processed by RNA-seq. RNA was harvested using Maxwell RSC miRNA Tissue Kit according to the manufacturer's protocol. Libraries for RNA-seq were prepared from 1 µg of total RNA using Truseq Total RNA Library Prep Kit (Illumina) following the manufacturer's protocols. Samples were sequenced paired-end 50 bp on an Illumina Novaseq6000 instrument.

Whole-exome sequencing. Sequencing was performed with the Agilent SureSelect All Exon v5 (experiments shown in Fig. 6c, d) or v7 (experiments shown in Fig. 6a, b), as per manufacturers' instructions. Libraries were sequenced with coverage >200X on a Novaseq 6000, with a PE 100 reads mode.

Multiseq—single-cell time-course for assaying drug tolerance

The single-cell time-course experiment for drug tolerance (shown in Fig. 7) was designed in order to process all time points at the same time. To do so, we performed an en-reverse experiment (i.e., starting from the last time point, d15). The same batch of cells was used for the entire experiment. Cells were passaged every 2 days and seeded either for passive culture or paclitaxel treatment (5×10^6 cells in 150 mm dishes). After 3 days of treatment (50 nM as described above), the medium was changed every 3 days without adding the drug. At day 15, all paclitaxel-treated time points were collected together, counted and processed with the Multiseq protocol[76]. Briefly, 500,000 cells for each sample (d5, d7, d9, d11, d13, d15) were resuspended in 180 µL of PBS and then labelled with 20 µL of a reaction mix composed of 2 µM of a sample-specific Barcode, 2 µM of LMO-Anchor (a kind gift of the Gartner Lab) and PBS. After 5′ of incubation in ice, we added 20 µl of reaction mix composed of 2 µM of co-anchor in PBS. After 5′ of incubation we stopped the reaction adding PBS/BSA 1%. We centrifuged and washed twice the cells before resuspending each sample in 125 µl PBS/BSA and pooling. After filtering, 25,000 cells were loaded on the Chromium Controller. Multi-seq library was isolated from the amplified cDNA and sequenced at 5000 barcode reads/cell depth.

RT-qPCR

Total RNA was extracted using Maxwell RSC miRNA Tissue Kit according to the manufacturer's protocol. One microgram of total RNAs was reverse-transcribed using ImProm-IITM Reverse Transcription System (Promega) and genes were analysed with Quantifast SYBR green master mix (Qiagen). *RPLP0* was used as a housekeeping gene. The complete list of primers used in this study is reported in Supplementary Data 18.

GBC library preparation from gDNA

The genomic DNA was extracted from 1 to 3 million cells (typically 3 million, coverage 300×) using the NucleoSpin Tissue kit (Macherey-Nagel). To enrich for GBCs, six parallel PCR reactions were performed in a final volume of 50 µl using 200 ng of genomic DNA, 0.2 µM of dNTPs mix, 0.5 µM of the following primers: F1: GGGTTTAAACGGGCCCTCTA and R4: GCCTGGAAGGCAGAAACGAC and amplified using Phusion™ High-Fidelity DNA Polymerase at the final concentration of 0.02 U/µl (Thermo-Scientific) (coverage 50×–500×), in its 5X Phusion HF buffer according to the following PCR protocol: (1) 98° for 30 s, (2) 98° for 7 s, then 60° for 25 s and 72° for 15 s (for 30 cycles), (3) 72° for 5 min.

At the end of the PCR, reactions were pooled and purified using QIaquick PCR purification kit (QIAGEN). Then, the eluted DNA samples were run on a 2% agarose gel and the 280 bp band was purified using the QIAquick Gel extraction kit (QIAGEN). Illumina libraries were generated from 10 ng of DNA, which was blunt-ended, phosphorylated, and tailed with a single 3′ A. An adapter with a single-base "T" overhang was added and the ligation products were purified and amplified to enrich for fragments that have adapters on both ends.

Libraries with distinct adapter indexes were multiplexed and sequenced (50 bp paired-end mode) on a Novaseq 6000 sequencer.

Single-cell library and GBC sublibrary preparation (scRNA or snRNA assays)

Single-cell suspensions (500–1000 cells/µl) were mixed with reverse transcription mix using the 10x Genomics Chromium Single-cell 3′ reagent kit protocol V2 (T0 MDA-MB-231 TGL and paclitaxel time-course exp2) or V3.1 (T1, paclitaxel time-course exp1) and loaded onto 10x Genomics Single-Cell 3′ Chips (www.10xgenomics.com). Libraries were generated as per manufacturers' instructions and sequenced on Illumina Novaseq 6000 Sequencing System (with a single- or dual-indexing format according to the manufacturer's protocol V2 or V3.1). We aimed at a coverage of 50 K reads/cell in each sequencing run. Multiome experiments were performed with the Chromium Single-Cell Multiome ATAC + Gene Expression Reagent Kits (V1). Nuclei suspensions (2000 nuclei/µL) were transposed and loaded onto Chromium Next GEM Chip J Single-Cell. Libraries were generated as per manufacturers' instructions and sequenced on Illumina NOVAseq 6000, aiming at 50 K RNA and 50 K ATAC reads/cell. To enrich for GBC reads, in a final volume of 50 µL, we amplified by PCR the Perturb-library cassette from 5 ng of the amplified cDNA (as in ref. 18) using 0.3 µM of dual-indices primers (forward: 5′-AATGATACGGCGACCACCGAGATCTACACCTCCAAGTTCACACTC TTTCCCTACACGACGCTCTTCCGATCT-3′; reverse: 5′-CAAGCAGAAGACGGCATACGAG ATCGAAGTATACGTGACTGGAGTTCAGACGTG TGCTCTTCCGATCTTAGCAAACTGGGGCACAAGC-3′) and amplified using Q5 2X master mix (M0541S, NEB) according to the following protocol: (1) 98° for 10 s, (2) 98° for 2 s, then 65° for 5 s and 72° for 10 s (25 cycles), (3) 72° for 1 min. The fragment band of the expected length (350–425 bp) was purified using EGEL 2% Power Snap Electrophoresis System (Thermo-Scientific) and checked at bioanalyzer before sequencing.

sgRNAs list

CRISPRi sgRNAs sequences targeting the two putative enhancers of *COL6A1* and *COL6A2* were designed using the web tool CRISPick (Broad Institute). The design region was defined by merging the ATAC module regions with the overlapping H3K27Ac signal from Encode (as shown in Fig. 5b) and exploiting the CRISPRi Range format for unstructured targeting provided by CRISPick. Inputs were: Enh_1 NC_000021.9:+:46048857-46050195 and Enh2 NC_000021.9:+:46052113-46053945. Selected sgRNAs were named according to the position relative to the start of the design window. We selected sgRNAs to cover the entire design window, choosing the sequences with higher on-target activity and filtering out those with potential off-target complementarity. As a negative control, we included three scrambled-sequence-sgRNAs for the CRISPRi experiments selected from the previous genome-wide CRISPRi screening library designed by the Weissman lab[77]. The sgRNAs chosen to target the promoters of COL6A1 and COL6A2 were also chosen from the same study. The list of sgRNAs sequence is reported in Supplementary Data 18.

Genetic barcode analysis

Genetic barcode calling. A GBC library is built for each sample separately, starting from the FASTQ files, in two steps. First, the 18nt-

long sequences located in the GBC locus are extracted using `seqkit amplicon` command from seqkit v2.1.0[78], with either 23–40nt- (DNA reads) or 12–29nt-long (cDNA reads) flanking regions and allowing one mismatch. A set S of sequences of length 18 is obtained and each $s \in S$ is assigned a weight $w(s)$, corresponding to its frequency in the FASTQ file. Note that, since the relative GBC abundance in a sample is unknown, not accounting for sequencing errors can result in an inflated estimate of sample clonality. Only sequencing errors in the form of mismatches are considered. The underlying assumption is that, if $s,s' \in S$ are sequenced from the same GBC species and carry d and d' mismatches, respectively, and that $d < d'$, then $w(s) \geq w(s')$. Thus, each GBC species i is associated with a subset $S_i \subseteq S$, where the true GBC sequence is the $s \in S_i$ with maximal weight. The frequency of i is the cumulative weight of the sequences in S_i. We infer the set of "true" GBC sequences and simultaneously correct for sequencing errors with the following procedure. Let $G = (V, E, w)$ be an undirected graph, where $V = S$, E is the set of edges connecting sequences at Hamming distance $\leq D$, and $w(s)$ is the frequency of s. We iteratively detect a collection of disjoint subgraphs $G_i = (S_i, E_i)$ induced by S_i, called *stars*, where a node $s \in S_i$ is the hub and all the other nodes are neighbours of s. Stars are defined according to the following conditions: (i) the hub s has maximum weight $w(s)$ in S_i, (ii) the cumulative weight across stars is maximal, and (iii) the fraction of neighbours of s in G that do not belong to S_i is $< f$, where $0 \leq f < 1$ is a fixed parameter. We compute a heuristic solution with a greedy approach. First, nodes are ordered by non-increasing weight. At each iteration, a new star is created, whose hub is the first node in the list and the other nodes are its neighbours, and the included nodes are removed from the list. The procedure ends when the first star that violates (iii) is found. We set $D = 1$ and $f = 0.2$. The final set of true GBC sequences is defined as the collection of detected hubs and their frequency is the cumulative weight across stars. We approximate the clone content of bulk DNA-Seq sequencing samples as the set of GBCs and their associated frequencies.

Definition of tumour-initiating clones and drug-tolerant clones. In bulk DNA-seq samples, we approximate clone content with GBC content and clone abundance with count per million reads (CPM). Clones whose frequency significantly differs between conditions are determined as follows. For each clone, we test if the CPM in the condition sample (treated or untreated) is significantly greater compared to the average across control samples (parental), using a Fisher's exact test. The universe is defined as the union of all clones across control and condition samples. P values are adjusted using the Bonferroni correction method. Clones with adjusted p value < 0.05 in at least $\lceil \frac{1}{2} n \rceil + 1$ condition samples are labelled *TICs* (if condition is "untreated tumour") or *DTCs* (if condition is "treated sample" or "treated tumour"), where n is the number of condition samples.

Single-cell RNA-seq data analysis (SUM159PT)

Cell barcode calling. A GBC reference is defined as the union of the GBC species obtained from cDNA reads, as described in the section "Genetic barcode calling", across all samples in the same experiment. Read alignment, UMI counting, cell-containing barcode (CB) calling, and GBC counting are performed using `cellranger count` from CellRanger v6.0 on the human reference genome GRCh38 v2020-A and the GBC reference. Parameter `--expect-cells` is set to 7500 (T0), 20,000 (T1 and exp1), and 3000 (exp2). Feature-CB count matrices are obtained, where features denote either genes or GBC species, and entries are UMI counts. Cellular barcodes with <5000 (T1 and exp1) or 10,000 (T0, exp2) gene UMIs are filtered out. Data post-processing is done with R[79].

MULTI-seq sample demultiplexing. Samples belonging to T1 and exp1 are demultiplexed using MULTI-seq barcodes (MBCs), as follows. MBC-containing reads are aligned to the reference barcode sequences using the R package deMULTIplex v1.0.2[76]. MBC-CB count matrices are obtained. Each MBC univocally identifies a sample, and a sample can be labelled with multiple MBCs. The automatic quantile-based thresholding procedure implemented in deMULTIplex fails to assign a unique MBC label to most of cellular barcodes in exp1, thus a custom demultiplexing procedure is used for both experiments. First, all MBCs with UMI count <2 are removed from every CB. For every cellular barcode i, an MBC j is marked as detected if $c_j \geq 15$, $c_j/c_{max} \geq 0.5$, and $p < 1e-5$, where c_j is the UMI count of j in i, c_{max} is the UMI count of the top abundant MBC in i, and p is the probability of observing a value equal or greater than c_j given a Poisson distribution with λ = average UMI count of j across cellular barcodes in the sample. Only cellular barcodes with exactly one detected MBC are retained and assigned to the corresponding sample.

Clone detection. Expressed GBCs in CBs are identified using the same procedure applied to MBCs and described in the section "MULTI-seq sample demultiplexing", using $c_j \geq 10$ and $c_j/c_{max} \geq 0.3$. A p value threshold is set only for samples d11–13–15 for exp1 ($p < 1e-5$) and sample d17 for exp2 ($p < 1e-10$). Note that GBC frequency at earlier time points is very low, hence the p value criterion has no effect. CBs can be assigned 0, 1, or >1 GBCs, the latter being an effect of multiple cell encapsulation (doublets) or multiple GBCs infecting the same cell (co-infection). We extract clone information from multi-GBC CBs by distinguishing co-infection and doublet events. High UMI doublets are removed using a sample-specific cutoff. Assuming that GBC expression is approximately constant across GBC species and across cells, differences in GBC UMI count are essentially due to droplet-specific mRNA capture. Moreover, the transcriptome size of a cell line model should be constant as well. We deduce that a single cell infected with k GBCs should display a higher GBC UMI fraction compared to that of multiple cells encapsulated in the same droplet that jointly account for k infection events. For simplicity, GBC species for multiple infection events within the same droplet are assumed to be pairwise distinct. The procedure works as follows. First, each CB set with k expressed GBCs is sorted by non-increasing values of c_i/u_i, where c_i and u_i are the GBC UMI count and the gene UMI count for CB i, respectively. We obtain a list of sets S_2, ..., S_m, where m is the maximum number of distinct GBC species expressed in a CB. Then, we iteratively classify the CBs of each set S_k separately, starting from the smallest k. A CB i is labelled as co-infection in two cases: (a) all GBCs expressed by i are also expressed in at least five CBs, including i, or (b) the "doublet probability" of i (i.e., the cumulative sample frequency of each single GBCs expressed in i) is below a sample-defined threshold. If neither (a) nor (b) hold, i is labelled as doublet. The procedure continues until the fraction of co-infection events in multi-GBC CBs is ≥1-D, where D is the expected doublet fraction in multi-GBC CBs. D is a sample-specific doublet rate estimate based on 10× guidelines and the number of called CBs. The clone pool of a single-cell sample is defined as the collection of single GBCs that occur in single infection and doublet events with $k = 2$ GBCs, plus all multi-GBC sets from co-infection events.

Clone comparison between single-cell and bulk samples. To assess the accuracy of clonality estimates from bulk samples, we compare the GBC species content between bulk DNA and single-cell RNA GBC sequencing samples from the same condition. We define the frequency of a GBC species in a bulk sample by CPM. Then, for a given pair (X_b, X_s) of bulk and single-cell samples, we define the value $y = f(x)$ as the fraction of GBC species with frequency $\geq x$ in X_b found as expressed in at least one cell of X_s. In practice, we compute $f(x)$ in steps of 20 CPM units. Consistent clone estimates in bulk and single-cell samples in a condition result in a monotonically non-decreasing trend of $f(\cdot)$. A cell c is classified as tumour-initiating or drug-tolerant (see section "Definition of tumour-initiating clones and drug-tolerant clones" above) if and

only if all GBCs expressed in c are tumour-initiating or drug-tolerant, respectively. Conversely, the single-cell cluster labelling is transferred to a bulk sample as follows: the GBC abundance (CPM) in the bulk sample is first normalised by its average abundance at baseline; then, the abundance of cluster C in the bulk sample is calculated as the contribution of all GBCs expressed by any cell in C.

Quality filtering and normalisation. All CBs with no expressed GBCs or classified as doublets are removed from subsequent analysis. Samples from technical replicates (same vial) are concatenated and processed using Seurat v4.0.5[80]. UMI counts are added a pseudocount = 1, divided by library size, multiplied by 10,000, and log-transformed (natural logarithm), to obtain log-normalised counts.

Dimensional reduction and clustering. Highly variable genes detection is performed on log-normalised counts using two different approaches: `min.var.plot` and `vst`. We set `xmin` = 0.1 and `xmax` = 10 for `min.var.plot`. The input parameters for each algorithm are let vary in a pre-defined set: dispersion.cutoff in {0.5, 1, 1, 5} for `min.var.plot`, and `nfeatures` in {1000, 2000, 5000} for `vst`. A cell-cycle score is computed using the Seurat function `CellCycleScoring` with default parameters. We observe a high cell-cycle effect in parental samples (T0 and T1); hence, the cell-cycle score computed as above is regressed out before scaling and centreing. PCA is performed on z-scores of the log-normalised counts on the reduced space of highly variable genes. A total of 50 PCs is computed. The optimal number n of PCs to retain for clustering is defined as the minimum n such that the standard deviation explained by the nth PC exceeds 50% of the average across the 40–50th PCs. Multiple Louvain clustering runs are performed on the selected PCs, for each highly variable gene set, by varying the number of neighbours k in the knn graph and the resolution parameter r in a pre-defined set: $k \in \{30, 40, 50\}$ and r from 0.1 to 0.8 in steps of 0.1. A second and third round of highly variable gene selection, PCA, and clustering are possibly performed after the removal of small clusters with very low UMI count, found in specific solutions, until the average UMI count is homogeneous across clusters. We redo the whole clustering procedure instead of just removing the small, low-quality clusters, because they are usually outliers in the expression profile of the sample and can affect the definition of highly variable genes. We obtain 9395 and 13,562 cells for T0 and T1, and 7884 and 10,698 cells for exp1 and exp2, respectively. For T0 and T1, (i) a silhouette score is computed for each clustering solution on the Euclidean distances calculated on the same PCs used for clustering and (ii) a ROGUE score[81] is computed for each clustering solution on the UMI counts, using default parameters, except for span =0.6. The clustering solution with the highest silhouette score is defined as optimal. To increase the number of detected clusters, we set a higher value for the resolution parameter, while maintaining the other parameters of the optimal solution unchanged (the highly variable gene set and the number of neighbours in the knn graph). The final clustering solutions for T0 and T1 are obtained with `vst` method by setting `nfeatures` = 1000 and clustering parameters $k = 40$ and $r = 0.5$, resulting in seven clusters for both experiments. Resolution 0.5 turned out to be a good compromise between silhouette width and ROGUE score on both T0 and T1. For exp1, we obtain three clusters using the following settings: `vst` method, `nfeatures` = 5000, $k = 30$, and $r = 0.1$. UMAP reduction[82] is run with RunUMAP using default parameters on the same input used for clustering.

Differential expression analysis. A differential expression analysis is run between conditions via `FindMarkers` function from Seurat with MAST method[83] accounting for sample ID as a covariate. We only test for genes detected (UMI count > 0) in at least 10% cells in either condition such that $|\log_2(FC)| \geq 0.25$. A gene is defined as differentially expressed if its adjusted p value is < 0.05 (Bonferroni correction) between the two conditions.

Computation of cell–cell distance distributions. To evaluate the relationship between clonality and gene expression at basal state, we compare sister cells (same clone) and non-sister cells (different clones) at T0 and T1. We note that a sister cell similarity measure within the same experiment would be biassed towards clones with vial-specific frequency >1, thus we opt for a comparison between experiments (1886 common clones). Cells of T0 and T1 are projected on a common latent space using canonical correlation analysis (CCA)[84], using the intersection of highly variable gene sets (defined as in the section "Dimensional reduction and clustering") between experiments. By definition, this integration approach removes dataset-specific complexity by keeping only shared correlations, hence the contribution of the experiment covariate to cell–cell similarity should be minimised compared to more conservative approaches. Inter-sample cell–cell Euclidean distances are computed on the CCA latent space on sister cell pairs and a random subset of non-sister pairs of the same size. A Gaussian kernel density estimation is computed for sister and non-sister distances separately to obtain a global distance distribution. Then, to evaluate the relationship between and within the DTC and the non-DTC pool before and after treatment, we computed the Euclidean distances between cells at T1 (baseline) and in exp1 (after treatment); finally, we evaluated the relationship between clonality and gene expression on the treated condition by computing intra-sample Euclidean distances at day ≥13 on the PC space of exp1 and exp2, as defined in the section "Dimensional reduction and clustering". We obtained Gaussian kernel density estimates for both cell–cell distance matrices as above.

Computation of clone sharedness and detection of cell subpopulations. To assess whether sister cell similarity is uniform across clones or is rather clone-specific, we introduce a *clone sharedness score* across clusters (see section "Dimensional reduction and clustering"). It is defined, for each pair of clusters i and j in T0 and T1, as $cs(i,j) = (n_{ij} \cdot n'_{ji})/(N_i \cdot N'_j)$, where n_{ij} (n'_{ji}) is the number of cells in i (in j) with at least one sister in j (in i) and N_i (N'_j) is the total number of cells in i (in j). Each cluster i in T0 is matched with the cluster j^{max} in T1 with maximum clone sharedness with i, namely, $j^{max} = \arg\max_j cs(i,j)$. The three clusters belonging to maximal clone sharedness pairs (i,j^{max}) are defined as subpopulations and labelled by non-increasing value of $cs(i,j^{max})$ as S1, S2, and S3. Transcriptionally stable clones are defined as the ones only found in one of the three subpopulations in both T0 and T1, they are 437 in total. This approach is independent of single-cell data integration, whose accuracy is difficult to assess when reliable markers are unknown[85,86]. We define the gene signature of $S \in \{S1, S2, S3\}$ as the set of genes that are differentially upregulated between S and its complement in both T0 and T1. We obtain 109, 29, and 27 genes for signatures S1, S2, and S3.

Definition of clone–clone pair propensity. We define a *pair propensity* score to measure the pairwise gene expression similarity bias between clones. For each cell, we consider its directed neighbourhood of size k, i.e., each cell has exactly k nearest neighbours and the neighbour relationship is not symmetric. Given two clones labels i and j, possibly identical, the observed (i,j) frequency f_{ij}^{obs} is the number of cell pairs (c_i,c_j) such that c_i and c_j belong to clones i and j, respectively, and c_j is a neighbour of c_i. The expected (i,j) frequency f_{ij}^{exp} is calculated based on the frequency of clones i and j and the neighbourhood size k, as follows. Given the neighbourhood of c_i, the probability that a given cell $c \neq c_i$ in the neighbourhood is labelled with j is given by $(n_j - I(i,j))/(N-1)$, where n_j is the number of cells labelled with j, N is the total number of cells in the sample, and $I(i,j)$ is the indicator function ($I(i,j) = 1$ if $i = j$ and $I(i,j) = 0$ if $i \neq j$). We obtain $f_{ij}^{exp} = n_i \cdot k \cdot (n_j - I(i,j))/(N-1)$, where n_i is the

number of cells labelled with clone i. The pair propensity of i and j is defined as $p_{ij} = f_{ij}^{obs}/f_{ij}^{exp}$. By iterating across all clone pairs, we obtain a non-symmetric $n \times n$ matrix P, where n is the number of distinct clone labels. The value of p_{ij} indicates the propensity of clones i and j to be closer ($p_{ij} > 1$) or farther ($p_{ij} < 1$) from each other in the sample than expected by chance, where $p_{ij} = 1$ denotes random association. Finally, we obtain a symmetric matrix $P' = (P^T P)^{\frac{1}{2}}$.

Detection of clone lineages. We use the above pair propensity definition to find groups of clones with mutually similar gene expression profiles on the treated condition at day ≥13. We first define the neighbour relationship, as follows. The top 1000 highly variable genes are computed with `vst` method on the union of the time points in exp1 and exp2, respectively. For each sample, a knn graph with $k = 40$ is built on the top n PCs (see section "Dimensional reduction and clustering"). Matrix P' (see section "Definition of clone–clone pair propensity") is computed on the top 20 highest frequency clones, clones i such that $p_{ii} < 1$ are discarded, and the matrix obtained is clustered using the `hclust` function with `ward.D2` method on Euclidean distances. We obtain two clusters A and B that serve as "anchor" to add the remaining clones. Clones not yet included in A nor in B are sorted by non-increasing frequency and we iteratively add one clone at a time, as follows. The first clone i in the ordering is considered and a pair propensity p_{iA} (p_{iB}) is computed between i and the union of clones in A (in B). If $p_{iB} < 1$ and $p_{iA} > p_{iB} + 1$, then i is added to A (likewise for B), otherwise it is discarded. The procedure continues until the last clone in the ordering is considered. Starting from the sets A and B computed on each of the four samples, we define lineage L1 (lineage L2) as the set of clones always assigned to A (to B) and detected at least once in both exp1 and exp2. We detect 11 clones for L1 and 17 clones for L2. We define the gene signature of L1 (of L2) as the set of genes that are differentially upregulated (downregulated) between cells in L1 and L2 in both exp1 and exp2. We obtain 42 and 46 genes for L1 and L2.

Functional annotation. We perform pathway enrichment analysis (REACTOME[87]) as follows. Genes with official gene symbols are first converted to ENTREZ identifiers, using limma v3.49.5[88] and clusterProfiler v4.1.4[89] packages, respectively. Enrichment significance is calculated with a Fisher's exact test using the `enrichPathway` function from clusterProfiler. Benjamini–Hochberg method is used to correct for multiple testing. We perform Gene Set Enrichment Analysis (GSEA[90]) as follows. Given a query case-control, the list of genes (universe) is ordered by non-increasing $\log_2$(FC) values; GSEA is run with `GSEA` function from clusterProfiler, with `nPerm` = 1000 sample permutations. MYC activity was computed with `ModuleScore` function from Seurat on T0 and exp1 cells using HALLMARK_MYC_TARGETS_V1 and HALLMARK_MYC_TARGETS_V2 (www.gsea-msigdb.org) as MYC target gene lists.

Single-cell RNA-seq data analysis (MDA-MB-231 TGL)

Read alignment, UMI counting, and CB calling are performed as for SUM159PT cells, with parameter `--expect-cells` set to 3000. CBs with < 200 detected genes (UMI > 0) and genes detected in < 3 cells are filtered out. Normalisation, scaling, cell-cycle regression, dimensional reduction, clustering and differential expression analysis are performed as described for SUM159PT cells. The final clustering solution is obtained with `vst` method by setting `nfeatures` = 1000 and clustering parameters $k = 40$ and $r = 0.5$, resulting in seven clusters.

Single-cell Multiome data analysis

Cell barcode calling. A GBC reference is defined as the union of the GBC sets computed from the scRNA-seq reads, as described in the section "Genetic barcode calling". Read alignment, UMI count, peak calling, and CB calling are performed using `cellranger-arc count` from CellRanger v6.0 on genome reference GRCh38 v2020-A-2.0.0. GBC counts are extracted using `cellranger count`, setting `--expect-cells` to 6200 (M0_1) and 8800 (M0_2) and `--include-introns`. CBs with < 10,000 UMIs are filtered out.

Clone detection and quality filtering. Expressed GBCs in CBs are identified using the same procedure applied to scRNA-seq samples and described in the section "Clone detection", using $c_j \geq 10$ and $c_j/c_{max} \geq 0.3$, and no p value threshold. All CBs with no expressed GBCs or classified as doublets are removed from subsequent analysis. We obtain 2446 and 2377 nuclei for M0_1 and M0_2, respectively.

Gene expression analysis. Cell clusters are defined using gene expression information only, on M0_1 and M0_2 separately. UMI count normalisation is calculated as for scRNA-seq datasets with Seurat v4.0.6. Highly variable genes are selected using `FindVariableFeatures` function from with method = `vst` and `nfeatures` = 5000. PCA is performed on the set of highly variable genes using `RunPCA` function with default parameters. The first 30 PCs are used to compute a Shared Nearest Neighbours graph via the `FindNeighbors` function, then `FindClusters` function is used to cluster the cells, using the SLM algorithm with resolution ranging between 0.2 and 1.2 in steps of 0.1. A silhouette width is calculated for each clustering solution on the Euclidean distance matrix computed on the same input used for clustering. Subpopulations S1, S2, and S3 in the two replicates are detected by matching the clusters to T0 and T1 using the sharedness score (see section "Computation of clone sharedness and detection of cell subpopulations"). The selected resolution value is the one that maximises the average silhouette width while keeping the three subpopulations distinct. We select resolution 0.4 and 0.6 for M0_1 and M0_2, respectively, and obtain 7 and 6 clusters. UMAP projection is calculated using `RunUMAP` with default parameters on the same input used for clustering. Cluster markers are extracted with the Wilcoxon rank-sum test implemented in `FindAllMarkers` function using default parameters. Differentially expresseSingle-cell RNA-seq data analysis (SUM159PT)". Subpopulation gene signatures are defined as for the scRNA-seq samples.

Chromatin accessibility analysis—data preprocessing. The region-cell count matrices, where regions are ATAC peaks and entries are fragment counts, are processed as follows. Raw fragment counts are normalised using term frequency-inverse document frequency (TF-IDF), which assigns higher importance to highly cell-specific regions. The active regions (fragment count ≥ 1) in at least ten cells are selected using the `FindTopFeatures` function from Signac v1.7.0[91]. Latent semantic indexing (LSI) is applied to reduce the dimensionality of the dataset. We compute 50 LSI dimensions. We keep dimensions from 2 to 50, as dimension 1 shows a positive correlation with sequencing depth. UMAP projection is calculated on the same LSI space, using default parameters.

We define pairs of regions associated with the same transposition event between replicates as those lying at a distance $\leq d$ on the genome. `findOverlap` function from GenomicRanges v1.46.1[92] is used to determine such pairs, by varying the gap size (i.e., the value of d) from −1000 (overlap) to 1000 (padding) in steps of 10. Depending on d, each region in M0_1 can have 0, 1 or multiple matching regions with M0_2, and vice versa. The idea is that a low d would fail to recognise regions stemming from the same transposition event, whereas a high d would result in many spurious matches. Low (high) values for d result in few (many) multiple matches. The value of d is chosen such that both u_d and $u_d/(N_d - u_d)$ are maximised, where u_d is the number of unique matches between M0_1 and M0_2 and N_d the total number of matches. We select $d = 0$ and obtain 120,414 and 120,377 matched regions in M0_1 and M0_2, respectively.

Chromatin accessibility analysis—topic modelling. We use a topic modelling approach to group regions that are consistently open in the same sets of cells while reducing data sparsity. Given a region-cell matrix, topic modelling outputs a set of *topics*, where each region has a probability of being assigned to a topic and each topic has a probability of being assigned to a cell. Modelling chromatin accessibility in this way has three advantages: first, the number of topics is typically orders of magnitude smaller compared to the number of regions; second, a cell's epigenome is expressed as the contribution of multiple topics; third, the importance of a region in a cell is interpretable, namely, as the combination of the weight of the region in a topic and the contribution of that topic in the cell's epigenome.

Specifically, we use the Latent Dirichlet allocation (LDA) model implemented in cisTopic v0.3.0[56] on each replicate separately, on the set of matching regions between the two replicates (both unique and multiple matches, see section "Chromatin accessibility analysis—data preprocessing"). The input to cisTopic is a raw region-cell count matrix, binarised by setting a cutoff at one fragment per region per cell. cisTopic is run using a total number of 1000 iterations and a burn-in period of 500 iterations to the Collapsed Gibbs Sampler. The procedure is repeated by varying the number of output topics n between 10 and 50 in steps of 10. We select the maximum log-likelihood solution and obtain $n = 40$ for both M0_1 and M0_2. Topics with Pearson correlation coefficient >0.5 between topic-cell probability and fragment count are discarded. We select 31 topics for M0_1 and 34 topics for M0_2.

Chromatin accessibility analysis—detection of ATAC modules. To select the topics that represent robust biological signals, we compute a topic reproducibility score between replicates, as follows. To this aim, an IDR[93] is calculated between every pair of topics in M0_1 and M0_2, respectively, on region-topic probabilities, using idr v2.0.3 tool with parameters `--peak-merge-method max --idr-threshold 0.05 --max-iterations 100`. Given topics t_1 and t_2 in M0_1 and M0_2 and a region r, the IDR statistic expresses the probability that the region-topic probabilities of r in t_1 and t_2 are different. We say that r is reproducible between t_1 and t_2 if IDR < 0.05. For each topic pair (t_1,t_2) in M0_1 and M0_2, we define the reproducibility score as the weighted mean between the number of reproducible regions and the 75th percentile of min {−125 $\log_2$(IDR), 1000} across those regions. We define the sets of regions with IDR < 0.05 in topic pairs with reproducibility score > 0 as *ATAC modules*. For each cell and for each module, we compute a *module AUC* (Area Under the Receiver Operating Characteristic curve) score, where the TF-IDF value is the predictor variable and the membership to the module is the response variable, using the `auc` function from pROC v1.18.0[94] and direction "<". The set of regions resulting from the union of all ATAC modules were annotated using the ENCODE registry of Regulatory Elements v2 (cCRE)[95] with a minimum overlap of 1 bp.

Relationship between epigenetic modules and subpopulations. We detect the ATAC modules explaining specific cell subpopulations as follows: for each ATAC module, we compute an AUC score where the predictor variable is the module AUC defined above and the response variable is either the subpopulation membership (S1, S2, S3) or the phenotype (TIC,DTC). Then, we compute Spearman's ρ between the module AUC for each of the top 40 reproducible ATAC modules and each gene among those detected in at least 20 cells in the highly variable gene pool (`vst` method, `nfeatures` = 5000) across all nuclei. We annotate each gene according to whether they belong to S1, S2, and S3 signatures (obtained from both scRNA-seq and Multiome gene expression data) or are annotated as human TF by Uniprot[96] (GO:0003700, taxon = human, gene product type = protein; 1435 TF in total). To identify putative cis-regulatory regions, we extract the genes in the transcriptional signatures of S1, S2, and S3 that are proximal to reproducible regions in epigenetic modules, with a neighbourhood size of ±50 kbp around the region, using `findOverlaps` function from GenomicRanges package.

Identification of enriched TFs. For a given module, we extract two sets of genes according to the following criteria: (i) Spearman's $\rho \geq 0.2$ between the module AUC and gene expression across nuclei (see above), or (ii) gene locus lying ≤100 kb away from any region in the module, as by GREAT 4.0.4[97] analysis, using the basal plus extension gene regulatory domain definition. These two sets are separately used as input to ChEA v.3[98] and the output ranked list of enriched TFs is obtained.

Whole-exome sequencing data analysis

Read alignment, variant calling, and extraction of reproducible CNVs. Reads were aligned with BWA (-t 16) v0.7.17[99] and CNVs were called with CNVkit v0.9.8[100] with default parameters, using the parental SUM159PT cells as a normal reference sample and providing the appropriate Agilent bed file (v5 or v7) as `target`. The chromosomal CNVs detected with CNVkit were retrieved. CNVs were intersected across replicates using `bedtools intersect` from bedtools v2.30.0[101], requiring ≥1 bp overlap. Only regions covered by all replicates are retained. A CNV consensus is then defined as the set of regions that span >80% of a CNV in each replicate.

Comparison with single-cell sequencing data. Regions from scATAC-seq assays of M0_1 and M0_2 are intersected with CNV consensus regions using subsetByOverlap function from GenomicRanges v1.45.0 with default parameters. Genes from scRNA-seq assays are intersected with CNV consensus regions using the same approach but allowing ±100 kbp around the CNV consensus regions. Gene loci coordinates are extracted from the annotation used in the single-cell Multiome analysis.

Bulk RNA-seq analysis of SUM159PT and tumours

Reads were trimmed, filtered and aligned using STAR v2.7.3[102]. Read count extraction and TPM normalisation were performed using `FeatureCounts`. The TM4SF1^{high} signature, consisting of 433 DEGs, was defined using edgeR (within Galaxy v3.36), considering the sorting batch as a factor, and selecting genes with $\log_2$(FC) > 1 and p value < 0.05 ($N = 3$ independent sorting batches). The functional analyses of TM4SF1^{high} DEGs were generated through the use of IPA (QIAGEN Inc.). Differential gene expression analysis in tumours vs 2D samples was performed with the Bioconductor Deseq2 package v1.34[103] using default parameters.

Analysis of gene signatures (S1, S2, S3) in bulk TCGA and METABRIC datasets

RNA expression data from primary bulk breast cancer patients were retrieved from Cbio Cancer Genomics Portal[104], selecting as studies METABRIC and TCGA and as Genomic Profiles "mRNA expression" (METABRIC: "mRNA expression z-scores relative to all samples (log microarray)"; TCGA: "mRNA expression z-scores relative to all samples (log RNA Seq V2 RSEM)"). Only complete samples were considered for this analysis. Molecular subtyping information was retrieved from the Cbio Portal. For each signature, the sum of the z-scores of all the expressed genes (genes expressed in <50% of the samples were excluded from the analysis) was calculated. Then, patients were stratified according to score quartiles (1st quartile, S1-high; 2nd and 3rd quartile, S1-mid; 4th quartile, S1-low; likewise for S2 and S3). The significance of the association with SUM159PT-derived signatures was calculated with a Chi-square test. Individual genes of the signatures were also analysed through multivariate linear correlation and visualised by a colour map of clustered correlations (K-means) using JMP17.2 (SAS).

Comparison with single-cell sequencing datasets from the literature

Triple-negative breast cancer (TNBC) scRNA-seq. Concatenated scRNA-seq data from primary TNBC samples were retrieved from GSE161529[50]. Filtered cell-count matrices were processed as in the section "Quality filtering and normalisation". Epithelial cells were defined by normalised EPCAM expression greater than the local minimum between the top two local maxima in the distribution. Cells with UMI count >20,000 were removed, resulting in 9063 epithelial cells for downstream analysis. Highly variable gene detection, PCA, and clustering are performed with `Seurat_preprocessing` function from Scissor 2.0.0[105]. We obtained nine clusters. Pseudo-bulk positive and negative phenotypes are generated for each subpopulation (S1, S2, S3) and its complement, respectively, on each scRNA-seq sample at baseline separately (T0_1, T0_2, T1_1, T1_2), resulting in $n = 8$ samples/phenotype pair (4 replicates/condition). Scissor is run on the single-cell dataset for each of the three phenotype pairs. Parameter `alpha` for the logit regression is let vary in {0.7, 0.8, 0.9}, until ≤15% phenotype-associated cells (both positive and negative) are detected (`cutoff` = 0.15). To quantify the enrichment of S1, S2, and S3 signatures across clusters, we computed the odds ratio for each signature–cluster pair. UMAP is run with default parameters on the top 10 PCs.

Curated Cancer Cell Atlas (3CA). We obtained the top 50 significant genes for each of the meta-programmes (MP) associated to malignant cells[51]. Gene symbols were first converted to synonyms to match the gene symbols in the gene-cell expression table, using `UpdateSymbolList` function from Seurat. We computed a joint gene expression for each gene list, using the `ModuleScore` function. Finally, we computed the area under the curve AUC(C,M), where the predictor variable is the ModuleScore associated to the meta-programme M and the response variable is true when a cell belongs to C, false otherwise.

Pancreatic ductal adenocarcinoma (PDAC) scRNA-seq. Association with subclonal dissemination for 2010 genes was retrieved from Simeonov et al.[20]. To map S1 and S3 signature genes, murine gene symbols were converted to human gene symbols with limma v3.49.5.

Lung adenocarcinoma scATAC-seq. Pre-metastatic gene scores for 20564 genes for Module 9 (RUNX2) were retrieved from LaFave et al.[52].

Statistics & reproducibility

No statistical method was used to predetermine the sample size, which was comparable to that reported in previous studies. Regarding single-cell experiments, the required number of individual single-cell profiles was determined to capture a sizable portion of the total number of GBCs, which was first assessed by bulk DNA sequencing. Statistical significance was measured as indicated in the figure legends. No data were excluded from the analysis performed in vivo. One single-cell time-course batch (MULTI-seq) was removed due to poor library quality. All experiments were replicated at least twice. The experiments were not randomised, as the study was performed on uniform biological material (i.e., a commercial cell line). For comparative in vivo experiments (e.g., TM4SF1 high vs bulk), animals were allocated randomly into the experimental groups. The investigators were not blinded to allocation during experiments and outcome assessment for the in vivo experiments (e.g., TM4SF1 high vs bulk).

Reporting summary

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

Data availability

The raw data generated in this study have been deposited in the ArrayExpress database under accession codes E-MTAB-13064, E-MTAB-13066 and E-MTAB-13896, in the GEO database under accession code GSE222596 and in the SRA database under accession code PRJNA922938. The processed data generated in this study have been deposited in Zenodo[106] and are provided in the Supplementary Information/Source Data file. The raw data used in this study are available in the GEO database under accession GSE161529. The processed data used in this study are available at https://doi.org/10.1016/j.ccell.2021.05.005 (https://ars.els-cdn.com/content/image/1-s2.0-S1535610821002713-mmc6.xlsx), https://doi.org/10.1038/s41586-023-06130-4 (https://www.dropbox.com/scl/fi/22xtcdh0z7bnn5g5ugz33/meta_programs_2023-07-13.xlsx?rlkey=2e7d718s46zybiyvjuptm67n4&dl=1) and https://doi.org/10.1016/j.ccell.2020.06.006 (https://www.cell.com/cms/10.1016/j.ccell.2020.06.006/attachment/f5a9ca73-3dc5-413d-99b0-24d348abf2f3/mmc4.xls). Source data are provided with this paper.

Code availability

The code used to reproduce the analysis reported in this study is available on github at https://github.com/nicassiolab/GBC_SUM159PT_paper and https://github.com/nicassiolab/GBC_SUM159PT_paper_figures. All codes have been deposited in Zenodo[107,108].

References

1. Marine, J. C., Dawson, S. J. & Dawson, M. A. Non-genetic mechanisms of therapeutic resistance in cancer. *Nat. Rev. Cancer* **20**, 743–756 (2020).
2. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer* **12**, 323–334 (2012).
3. Visvader, J. E. & Lindeman, G. J. Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. *Nat. Rev. Cancer* **8**, 755–768 (2008).
4. Lapidot, T. et al. A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645–648 (1994).
5. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat. Med.* **3**, 730–737 (1997).
6. Ricci-Vitiani, L. et al. Identification and expansion of human colon-cancer-initiating cells. *Nature* **445**, 111–115 (2007).
7. Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl Acad. Sci. USA* **100**, 3983–3988 (2003).
8. Batlle, E. & Clevers, H. Cancer stem cells revisited. *Nat. Med.* **23**, 1124–1134 (2017).
9. Singh, S. K. et al. Identification of human brain tumour initiating cells. *Nature* **432**, 396–401 (2004).
10. Lathia, J. D., Mack, S. C., Mulkearns-Hubert, E. E., Valentim, C. L. & Rich, J. N. Cancer stem cells in glioblastoma. *Genes Dev.* **29**, 1203–1217 (2015).
11. Basile, K. J. & Aplin, A. E. Resistance to chemotherapy: short-term drug tolerance and stem cell-like subpopulations. *Adv. Pharm.* **65**, 315–334 (2012).
12. McCarthy, D. J. et al. Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes. *Nat. Methods* **17**, 414–421 (2020).
13. Ross, E. M. & Markowetz, F. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol.* **17**, 69 (2016).
14. Satas, G., Zaccaria, S., Mon, G. & Raphael, B. J. SCARLET: single-cell tumor phylogeny inference with copy-number constrained mutation losses. *Cell Syst.* **10**, 323–332 e328 (2020).
15. Zhou, Z., Xu, B., Minn, A. & Zhang, N. R. DENDRO: genetic heterogeneity profiling and subclone detection by single-cell RNA sequencing. *Genome Biol.* **21**, 10 (2020).
16. Biddy, B. A. et al. Single-cell mapping of lineage and identity in direct reprogramming. *Nature* **564**, 219–224 (2018).

17. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, eaaw3381 (2020).
18. Dixit, A. et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 e1817 (2016).
19. Jindal, K., VanHorn, S. & Morris, S. A. New dual-channel system records lineage in high definition. *Nat. Methods* **19**, 38–39 (2022).
20. Simeonov, K. P. et al. Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states. *Cancer Cell* **39**, 1150–1162 e1159 (2021).
21. Gutierrez, C. et al. Multifunctional barcoding with ClonMapper enables high-resolution study of clonal dynamics during tumor evolution and treatment. *Nat. Cancer* **2**, 758–772 (2021).
22. Oren, Y. et al. Cycling cancer persister cells arise from lineages with distinct programs. *Nature* **596**, 576–582 (2021).
23. Echeverria, G. V. et al. High-resolution clonal mapping of multi-organ metastasis in triple negative breast cancer. *Nat. Commun.* **9**, 5079 (2018).
24. Merino, D. et al. Barcoding reveals complex clonal behavior in patient-derived xenografts of metastatic triple negative breast cancer. *Nat. Commun.* **10**, 766 (2019).
25. Nguyen, L. V. et al. DNA barcoding reveals diverse growth kinetics of human breast tumour subclones in serially passaged xenografts. *Nat. Commun.* **5**, 5871 (2014).
26. Karras, P. et al. A cellular hierarchy in melanoma uncouples growth and metastasis. *Nature* **610**, 190–198 (2022).
27. Yang, D. et al. Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. *Cell* **185**, 1905–1923 e1925 (2022).
28. Black, J. R. M. & McGranahan, N. Genetic and non-genetic clonal diversity in cancer evolution. *Nat. Rev. Cancer* **21**, 379–392 (2021).
29. Salgia, R. & Kulkarni, P. The genetic/non-genetic duality of drug 'resistance' in cancer. *Trends Cancer* **4**, 110–118 (2018).
30. Nam, A. S., Chaligne, R. & Landau, D. A. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. *Nat. Rev. Genet.* **22**, 3–18 (2021).
31. Prat, A. et al. Characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic molecular subtypes. *Breast Cancer Res. Treat.* **142**, 237–255 (2013).
32. Saunus, J. M. et al. Multidimensional phenotyping of breast cancer cell lines to guide preclinical research. *Breast Cancer Res. Treat.* **167**, 289–301 (2018).
33. Gupta, P. B. et al. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* **146**, 633–644 (2011).
34. Bierie, B. et al. Integrin-beta4 identifies cancer stem cell-enriched populations of partially mesenchymal carcinoma cells. *Proc. Natl Acad. Sci. USA* **114**, E2337–E2346 (2017).
35. Watson, A. W. et al. Breast tumor stiffness instructs bone metastasis via maintenance of mechanical conditioning. *Cell Rep.* **35**, 109293 (2021).
36. Fei, F. et al. Role of metastasis-induced protein S100A4 in human non-tumor pathophysiologies. *Cell Biosci.* **7**, 64 (2017).
37. Low, R. R. J. et al. S100 family proteins are linked to organoid morphology and EMT in pancreatic cancer. *Cell Death Differ.* **30**, 1155–1165 (2023).
38. Chen, J. et al. Transmembrane 4L six family member 1 suppresses hormone receptor-positive, HER2-negative breast cancer cell proliferation. *Front. Pharmacol.* **13**, 770993 (2022).
39. Hou, S. et al. TM4SF1 promotes esophageal squamous cell carcinoma metastasis by interacting with integrin alpha6. *Cell Death Dis.* **13**, 609 (2022).
40. Xing, P. et al. Upregulation of transmembrane 4 L6 family member 1 predicts poor prognosis in invasive breast cancer: a STROBE-compliant article. *Medicine* **96**, e9476 (2017).
41. Yang, J. C. et al. TM4SF1 promotes metastasis of pancreatic cancer via regulating the expression of DDR1. *Sci. Rep.* **7**, 45895 (2017).
42. Ferrari, E. & Gandellini, P. Unveiling the ups and downs of miR-205 in physiology and cancer: transcriptional and post-transcriptional mechanisms. *Cell Death Dis.* **11**, 980 (2020).
43. Dong, M., Dong, Z., Zhu, X., Zhang, Y. & Song, L. Long non-coding RNA MIR205HG regulates KRT17 and tumor processes in cervical cancer via interaction with SRSF1. *Exp. Mol. Pathol.* **111**, 104322 (2019).
44. Liu, L., Li, Y., Zhang, R., Li, C., Xiong, J. & Wei, Y. MIR205HG acts as a ceRNA to expedite cell proliferation and progression in lung squamous cell carcinoma via targeting miR-299-3p/MAP3K2 axis. *BMC Pulm. Med.* **20**, 163 (2020).
45. Mendez, O. et al. Extracellular HMGA1 promotes tumor invasion and metastasis in triple-negative breast cancer. *Clin. Cancer Res.* **24**, 6367–6382 (2018).
46. Alborghetti, M. R., Furlan, A. S. & Kobarg, J. FEZ2 has acquired additional protein interaction partners relative to FEZ1: functional and evolutionary implications. *PLoS ONE* **6**, e17426 (2011).
47. Zhang, X. et al. Identification of ribosomal protein S25 (RPS25)-MDM2-p53 regulatory feedback loop. *Oncogene* **32**, 2782–2791 (2013).
48. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
49. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
50. Pal, B. et al. A single-cell RNA expression atlas of normal, pre-neoplastic and tumorigenic states in the human breast. *EMBO J.* **40**, e107333 (2021).
51. Gavish, A. et al. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature* **618**, 598–606 (2023).
52. LaFave, L. M. et al. Epigenomic state transitions characterize tumor progression in mouse lung adenocarcinoma. *Cancer Cell* **38**, 212–228 e213 (2020).
53. Chen, G. et al. Targeting TM4SF1 exhibits therapeutic potential via inhibition of cancer stem cells. *Signal Transduct. Target. Ther.* **7**, 350 (2022).
54. Gao, H. et al. Multi-organ site metastatic reactivation mediated by non-canonical discoidin domain receptor 1 signaling. *Cell* **166**, 47–62 (2016).
55. Weaver, B. A. How Taxol/paclitaxel kills cancer cells. *Mol. Biol. Cell* **25**, 2677–2681 (2014).
56. Bravo Gonzalez-Blas, C. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).
57. Lachmann, A. et al. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).
58. Ansieau, S., Morel, A. P., Hinkal, G., Bastid, J. & Puisieux, A. TWISTing an embryonic transcription factor into an oncoprotein. *Oncogene* **29**, 3173–3184 (2010).
59. Beck, B. et al. Different levels of Twist1 regulate skin tumor initiation, stemness, and progression. *Cell Stem Cell* **16**, 67–79 (2015).
60. Nobre, A. R. et al. ZFP281 drives a mesenchymal-like dormancy program in early disseminated breast cancer cells that prevents metastatic outgrowth in the lung. *Nat. Cancer* **3**, 1165–1180 (2022).
61. Brown, M. S. et al. Phenotypic heterogeneity driven by plasticity of the intermediate EMT state governs disease progression and metastasis in breast cancer. *Sci. Adv.* **8**, eabj8002 (2022).
62. Gilbert, L. A. et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).

63. Larson, M. H. et al. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* **8**, 2180–2196 (2013).
64. Loukas, I. et al. Selective advantage of epigenetically disrupted cancer cells via phenotypic inertia. *Cancer Cell* **41**, 70–87 e14 (2023).
65. Masiulionyte, B., Valiulyte, I., Tamasauskas, A. & Skiriute, D. Metallothionein genes are highly expressed in malignant astrocytomas and associated with patient survival. *Sci. Rep.* **9**, 5406 (2019).
66. Wang, X., Yan, J., Shen, B. & Wei, G. Integrated chromatin accessibility and transcriptome landscapes of doxorubicin-resistant breast cancer cells. *Front. Cell Dev. Biol.* **9**, 708066 (2021).
67. Cho, S. W. et al. Promoter of lncRNA gene PVT1 is a tumor-suppressor DNA boundary element. *Cell* **173**, 1398–1412 e1322 (2018).
68. Tseng, Y. Y. et al. PVT1 dependence in cancer with MYC copy-number increase. *Nature* **512**, 82–86 (2014).
69. Dhimolea, E. et al. An embryonic diapause-like adaptation with suppressed Myc activity enables tumor treatment persistence. *Cancer Cell* **39**, 240–256 e211 (2021).
70. Zhou, M. et al. MicroRNA-125b confers the resistance of breast cancer cells to paclitaxel through suppression of pro-apoptotic Bcl-2 antagonist killer 1 (Bak1) expression. *J. Biol. Chem.* **285**, 21496–21507 (2010).
71. Lu, Y. et al. lncRNA MIR100HG-derived miR-100 and miR-125b mediate cetuximab resistance via Wnt/beta-catenin signaling. *Nat. Med.* **23**, 1331–1341 (2017).
72. Dave, B., Mittal, V., Tan, N. M. & Chang, J. C. Epithelial-mesenchymal transition, cancer stem cells and treatment resistance. *Breast Cancer Res.* **14**, 202 (2012).
73. Roche, J. The epithelial-to-mesenchymal transition in cancer. *Cancers* **10**, 52 (2018).
74. Pece, S. et al. Biological and molecular heterogeneity of breast cancers correlates with their cancer stem cell content. *Cell* **140**, 62–73 (2010).
75. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882 e1821 (2016).
76. McGinnis, C. S. et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).
77. Horlbeck, M. A. et al. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife* **5**, e19760 (2016).
78. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* **11**, e0163962 (2016).
79. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2022).
80. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 e3529 (2021).
81. Liu, B. et al. An entropy-based metric for assessing the purity of single cell populations. *Nat. Commun.* **11**, 3155 (2020).
82. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
83. 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).
84. 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).
85. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
86. Tran, H. T. N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
87. Gillespie, M. et al. The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–D692 (2022).
88. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
89. Wu, T. et al. ClusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
90. Sergushichev, A. A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. Preprint at *bioRxiv* https://doi.org/10.1101/060012 (2016).
91. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
92. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
93. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011). 1728.
94. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **12**, 77 (2011).
95. Consortium, E. P. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
96. Consortium, T. U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2018).
97. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
98. Keenan, A. B. et al. ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res.* **47**, W212–W224 (2019).
99. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
100. Talevich, 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).
101. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
102. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
103. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
104. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
105. Sun, D. et al. Identifying phenotype-associated subpopulations by integrating bulk and single-cell sequencing data. *Nat. Biotechnol.* **40**, 527–538 (2022).
106. Nadalin, F. Multi-omic lineage tracing predicts the transcriptional, epigenetic and genetic determinants of cancer evolution—processed data- https://doi.org/10.5281/zenodo.10912157 (2023).
107. Nadalin, F. Multi-omic lineage tracing predicts the transcriptional, epigenetic and genetic determinants of cancer evolution—reproducibility https://doi.org/10.5281/zenodo.10979191 (2023).
108. Nadalin, F. Multi-omic lineage tracing predicts the transcriptional, epigenetic and genetic determinants of cancer evolution—source code- https://doi.org/10.5281/zenodo.10979121 (2023).

Acknowledgements

This work was supported by grants from the Associazione Italiana per la Ricerca sul Cancro (AIRC) to F. Nicassio (IG18774 and IG22851), from the Fondazione Cariplo to F. Nicassio (2015-0590) and M.J.M. (2016-0615),

and from "National Center for Gene Therapy and Drugs based on RNA Technology" (CN00000041) supported by European Union—NextGenerationEU PNRR MUR—M4C2 to F. Nicassio; and "Roche per la ricerca 2018" to M.J.M. F. Nadalin was supported by a REBIT-POD fellowship. B.G. was supported by a FIRC-AIRC fellowship for Italy (22438). J.C.M. and I.P. acknowledge funding from EMBL member states. M.P.P. is a PhD student within the European School of Molecular Medicine (SEMM). Figures 1a, b; 2a, h; 3a; 4a; 5c; 6d, e; 7a and Supplementary Figs. 1a, 5a were created with BioRender.com and released under a Creative Commons Attribution-NonCommercial-NoDerivs (CC-BY-NC-ND) 4.0 International license. We thank Chiara Tordonato for help with mice experiments; Leah Rosen and Magdalena Strauss for input on barcode analysis; Pier Giuseppe Pelicci, Niccolò Roda and Valentina Gambino for help with the Perturb-seq lentiviral infection. We acknowledge support by the technological units at the European Institute of Oncology (IEO), in particular to the Genomic Units and Luca Rotta, the Sorting Service and Simona Ronzoni, the tissue culture facility, the imaging unit and the bioinformatics unit; the EMBL-EBI gene expression team; the Mouse Facility and the DNA sequencing service at Cogentech. We thank Pier Giuseppe Pelicci and all the participants to the Single-Cell Technoshot (1.0, 2.0) by IEO for support and discussion; Alvis Brazma for insightful discussions throughout the project; Marioni group, Papatheodorou group, Nicassio group for discussions; Marco Cosentino Lagomarsino and Dafne Di Campigli di Giammartino for critical reading the manuscript; Nancy George for proofreading.

Author contributions

F. Nadalin: conceptualisation, methodology, investigation, writing—original draft preparation. M.J.M.: conceptualisation, investigation, writing—review and editing. M.P.P.: investigation, writing—review & editing. P.F.: investigation, writing—review & editing. S.P.: formal analysis. M.C.: investigation. P.B.: investigation. C.R.: resources. B.G.: resources. I.P.: supervision, writing—review & editing. J.C.M.: supervision, writing—review & editing. F. Nicassio: conceptualisation, funding acquisition, supervision, writing—original draft preparation.

Competing interests

J.C.M. has been an employee of Genentech since September 2022. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-51424-4.

Correspondence and requests for materials should be addressed to F. Nadalin or F. Nicassio.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.