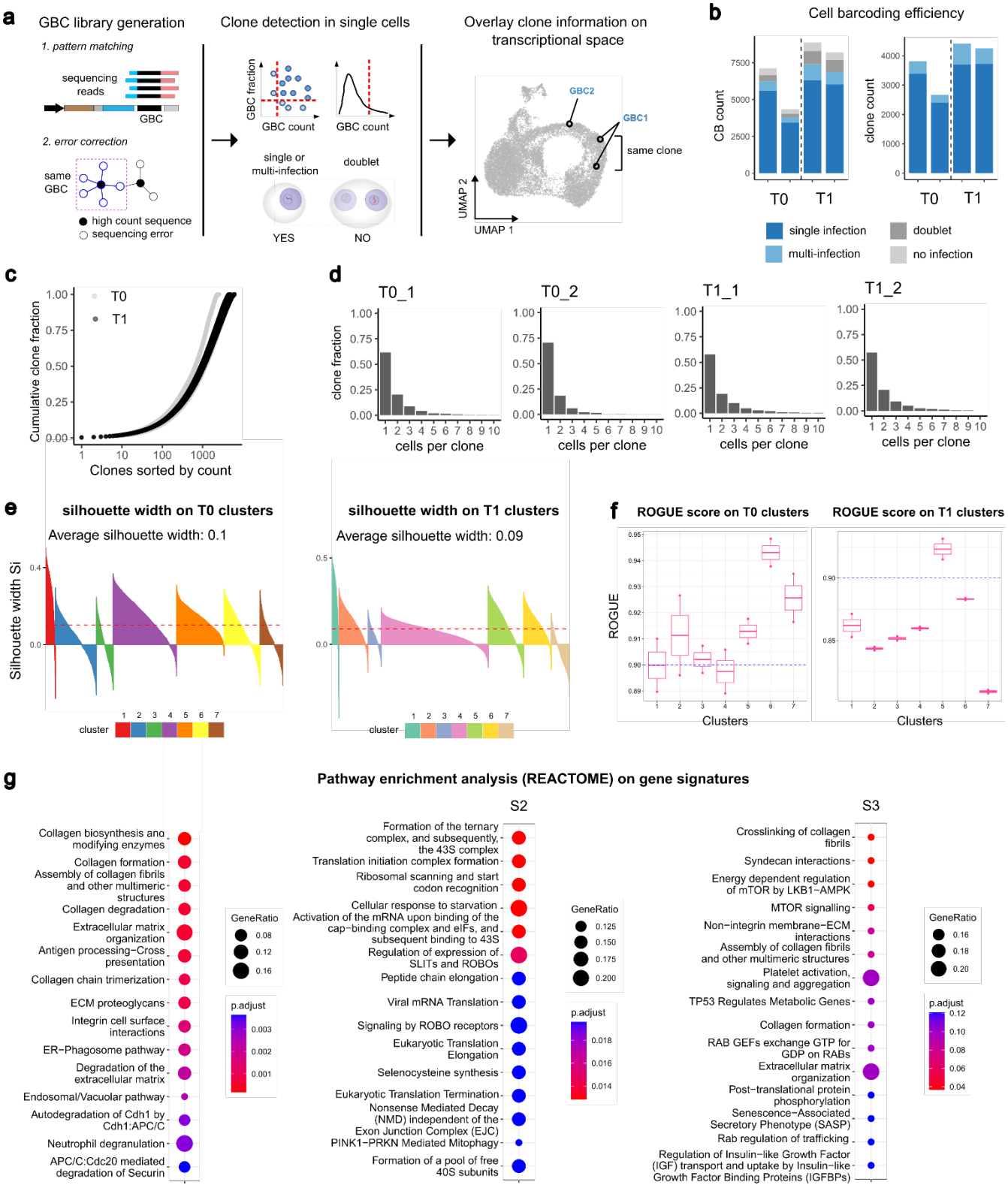
Supplementary Information - Multi-omic lineage tracing predicts the transcriptional, epigenetic and genetic determinants of cancer evolution

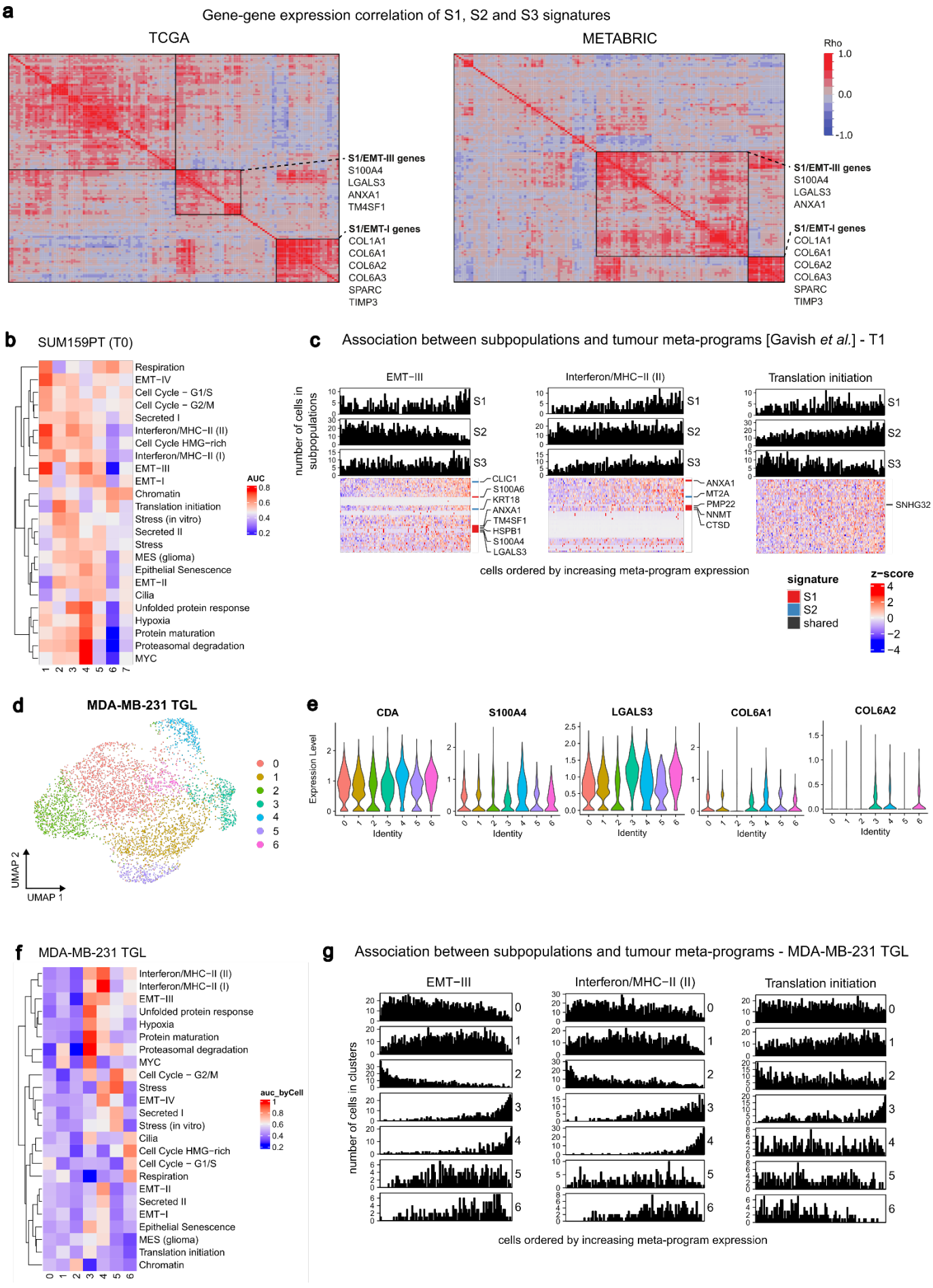
Table of contents

1. Supplementary Figures



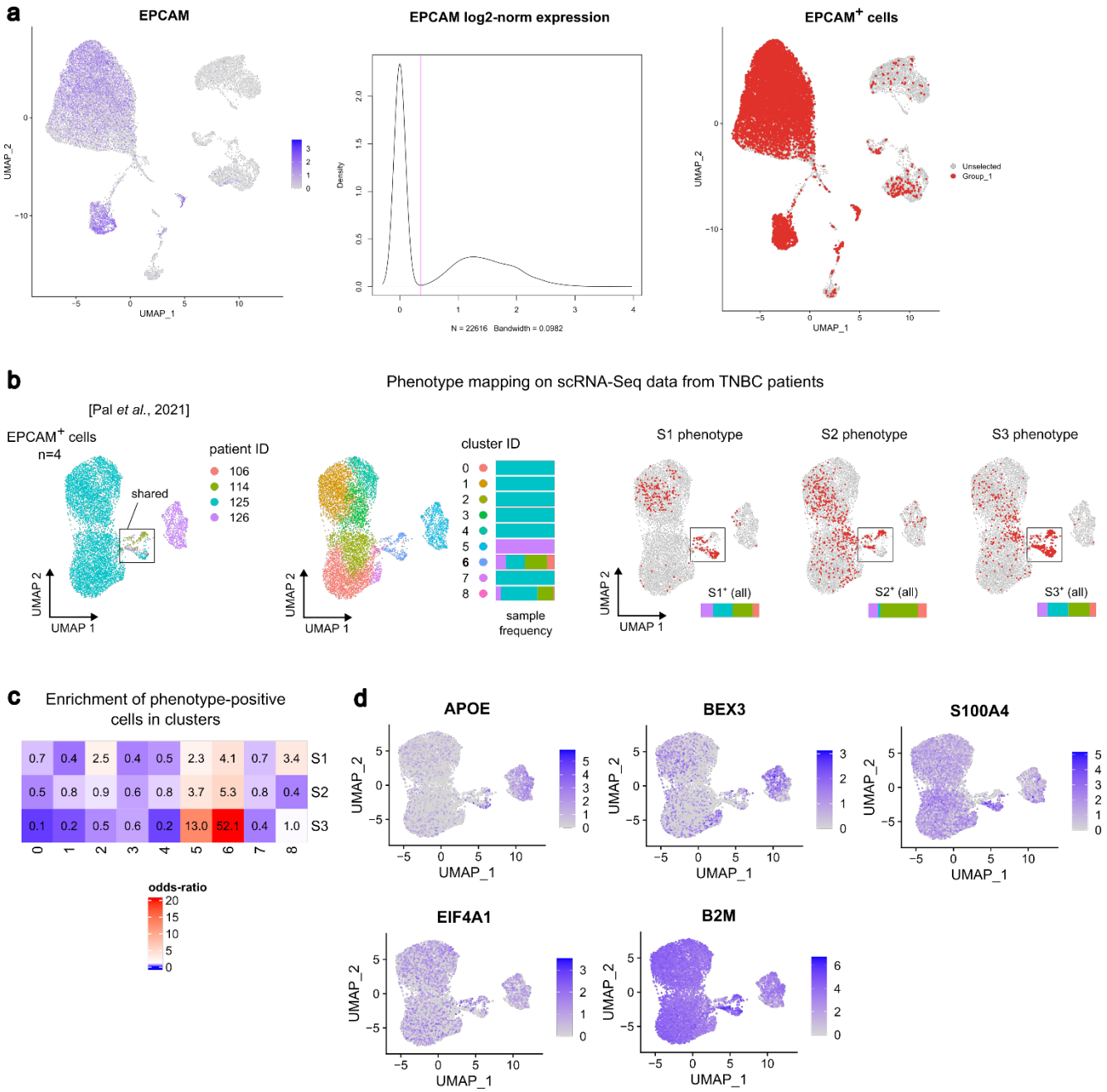
Supplementary Figure 1. Lineage tracing identifies TNBC subpopulations. a. Left: the GBC sequence library is generated in two analysis steps: 1. sequences on the GBC locus are extracted from the reads via pattern matching; 2. error-correction is performed using a graph approach. Centre: expressed GBCs in single cells are detected, single-cell-containing droplets are selected, and cells are assigned to clones accordingly. Right: sample UMAP representation showing the overlay of clone information on cells in gene expression space. **b.** Clone calling for T0 and T1 (n=2 replicates per condition). Left: cellular barcode (CB) classification into single infection, multi-infection, doublet, and no infection; CB count (left) and clone count (right) are shown. Clones from both single-infection, co-infections, and doublets with exactly 2 expressed GBCs are retained (see Methods). **c.** Comparison of cumulative

clone distributions in parental (T0) and untreated (T1) samples. **d.** Relative clone frequency as a function of the number of cells per clone for T0 and T1 in each duplicate. **e.** Silhouette width for each cell in the clustering solutions shown on Figure 1d, for T0 (left) and T1 (right) scRNA-Seq samples (n=2 per time point). **f.** ROGUE score for each cell in the clustering solutions shown on Figure 1d, for T0 (left) and T1 (right) scRNA-Seq samples (n=2 per time point). **g.** Pathway enrichment analysis (REACTOME) of subpopulation gene signatures. The top 15 significantly enriched terms (q-value < 0.1) are reported and sorted by non-increasing q-value. The size of the circles is proportional to the fraction of genes found in each pathway for each signature. Source data are provided as a Source Data file. [a. created with Biorender.com released under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license].

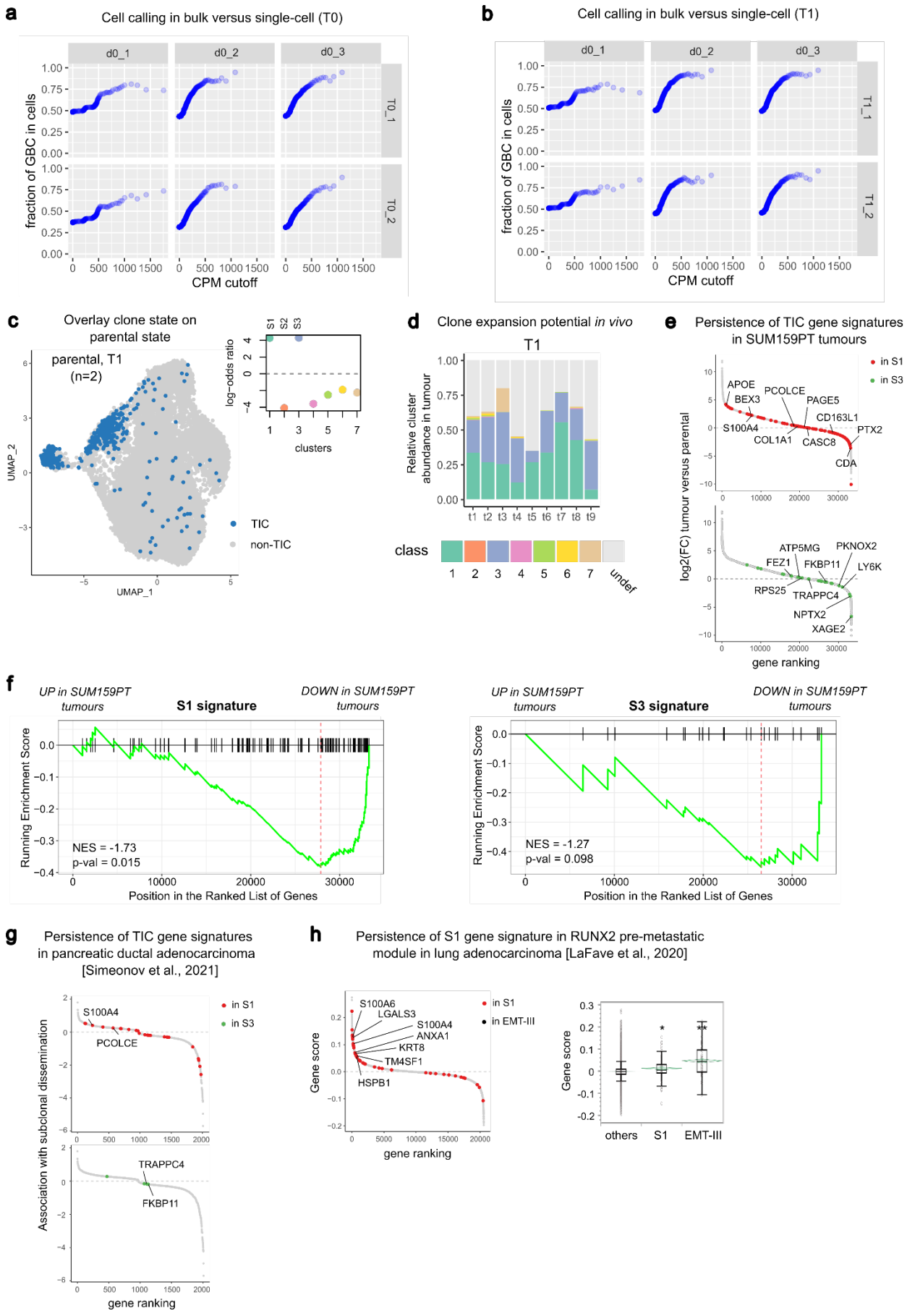


Supplementary Figure 2. Characterization of S1 and S3 subpopulations in cancer datasets. a. Clustering (k-means) of linear correlations between the expression level (Z-Score, see Methods) of each gene belonging to S1, S2

and S3 gene signatures and TCGA (left) or METABRIC (right) datasets. b. Association of gene meta-programs from Gavish et al. [1], with clusters at T0 (numbered as in Figure 1d). Only "shared" meta-programs are reported. The entries are the AUC values where the predictor is the aggregate meta-program expression, and the response is the membership to each cluster. **c**. Detailed association for three meta-programs highly associated with S1, S1-S3, and S2 at T1. The columns of the heatmaps represent the cells ordered by non-decreasing meta-program expression; the genes common to the subpopulation signatures are marked in color and labelled. The bar plots show the binned cell count (100 bins) for each subpopulation. **d.** UMAP representation of MDA-MB-231 TGL (4610 cells) coloured by cluster. **e.** Violin plots showing the log-normalised expression of selected S1 signature genes across MDA-MB-231 TGL clusters. **f.** Same as b., for MDA-MB-231-TGL. **g.** Bar plots as in c., where bins are defined on MDA-MB-231 TGL cells and cell count is computed for each cluster. Source data are provided as a Source Data file.

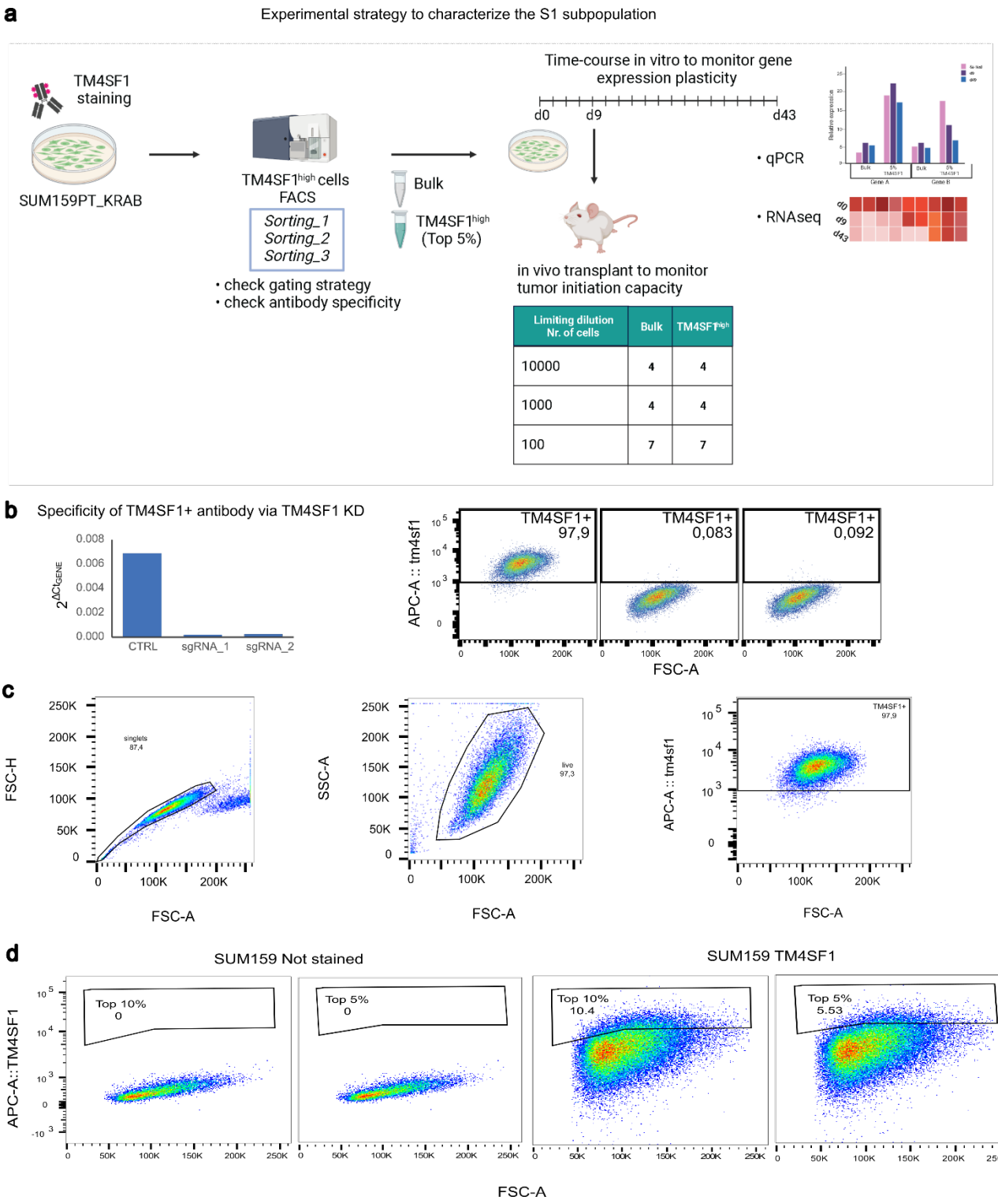


Supplementary Figure 3. Mapping subpopulations on scRNA-Seq data from TNBC patients. a. Selection of epithelial cells from Pal et al. dataset [2]. Left: cells on UMAP gene expression space from n=4 primary TNBC samples, coloured according to log-normalised EPCAM expression. Centre: gaussian kernel density of log-normalised EPCAM expression; a vertical line coloured in magenta separates non-epithelial (EPCAM-, left mode) and epithelial cells (EPCAM+, right mode). Right: cells on UMAP gene expression space where epithelial cells are coloured in red and non-epithelial cells are coloured in grey. **b.** Transcriptional phenotype inference for S1, S2, and S3 subpopulations on Pal et al. dataset [2]. Left: the epithelial (EPCAM+) cells from n=4 primary TNBC samples, defined as in A., are plotted on gene expression space (UMAP) and coloured either by sample or by cluster (9063 cells in total); the composition of the whole set of cells and of the cluster of cells shared among samples (cluster 6, indicated with a rectangle) is reported with a coloured bar. Right: Scissor output for each phenotype, as defined by bulk S1, S2, and S3 gene expression, on the EPCAM+ cells defined above; cells predicted as phenotype-positive are highlighted in red; the sample composition of phenotype-positive cells is reported with a coloured bar. **c.** Odds-ratio comparing the fraction of cells in each cluster-phenotype pair; clusters (in columns) and phenotype-positive cell subsets (in rows) as in b. **d.** Epithelial cells plotted on gene expression space (UMAP), as in b., coloured according to the log-normalised expression of S1 signature genes; the reported genes are significantly upregulated both in S1+ cells (compared to S1 complement) and in SUM159PT tumours (compared to baseline SUM159PT expression).



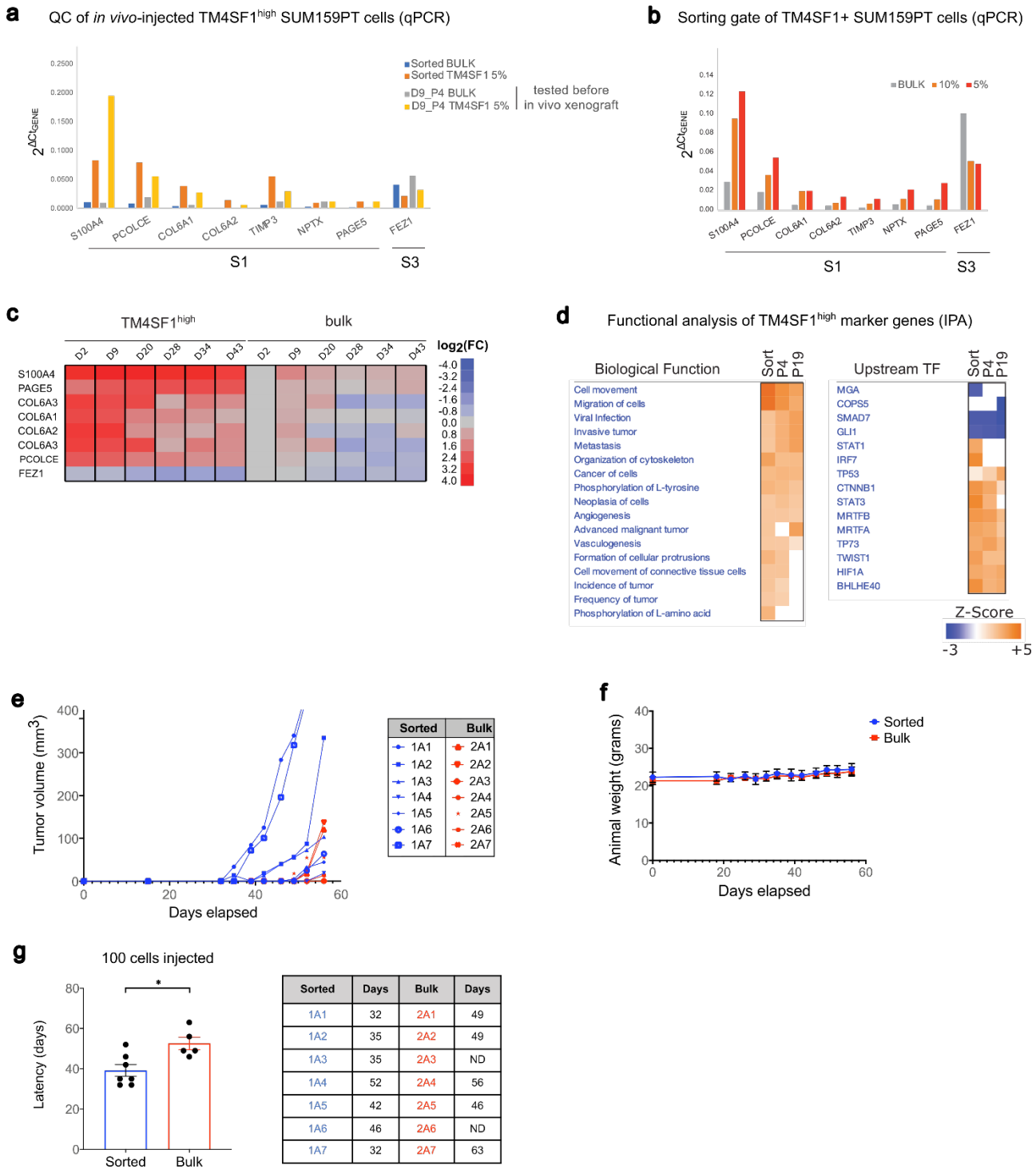
Supplementary Figure 4. Association of tumour initiating clones with TNBC subpopulations. a. Clone estimate from bulk DNA-seq (CPM, T0, *x* axis) versus clone calling from scRNA-Seq (GBC fraction, T0, *y* axis). The x axis represents the midpoint of CPM bins (see Methods). **b.** Clone estimate from bulk DNA-seq (CPM, T0, *x* axis) versus clone calling from scRNA-Seq (GBC fraction, T1, *y* axis). **c.** Mapping of TICs at parental state (T1). Left:

UMAP representation of T1 cells on gene expression space (477 cells, coloured in blue). Right: log-odds-ratio comparing cluster assignment and TIC labelling at T1. **d.** Association between parental state (T1) and clone expansion *in vivo*. The bar plot shows the relative abundance of T1 clusters in every tumour (unassigned clones shown in grey). **e.** Differential expression analysis in SUM159PT tumours versus the parental population. Genes are ranked by non-decreasing $\log_2$(FC) expression between tumours (n=5) and parental samples (n=6); genes in the S1 (S3) signature are marked in red (green). Genes with top $\log_2$(FC) in the scRNA-Seq assay are labelled. **f.** GSEA of S1 (left) and S3 (right) gene signatures on the list of genes ranked by $\log_2$(FC) expression in SUM159PT tumours compared to the baseline. **g.** S1 and S3 gene signatures in the scRNA-Seq dataset of pancreatic ductal adenocarcinoma from Simeonov et al.[3] Genes are ranked by non-decreasing association with subclonal dissemination in pancreatic metastatic clones. Genes in S1 and S3 signatures are coloured as in e. **h.** Gene signatures in the scATAC-Seq dataset of lung adenocarcinoma from LaFave et al.[4] Left: genes are ranked by gene score; S1 signature genes are coloured in red, with those that are also found among the top 50 of the EMT-III meta-program, as defined in Gavish et al.,[1] labelled. Right: boxplot (with centre: median, bound of box: upper and lower quartile, whiskers: 1.5·IQR) of the gene scores across different groups. Also shown are individual points (grey circles) and the mean diamonds (green line); * p-value = 0.001, ** p-value < 0.0001, Wilcoxon rank-sum test). Source data are provided as a Source Data file.



Supplementary Figure 5. Isolation of S1 subpopulation by TM4SF1 expression. a. Experimental design to characterise the S1 subpopulation: SUM159PT (expressing a doxycycline inducible dCas9-KRAB transgene, SUM159PT_KRAB hereafter) were stained with an APC-conjugated TM4SF1 antibody. Bulk and TM4SF1^{high} subpopulations were FAC-sorted (n=3) and processed for RNA extraction and subsequent RT-qPCR and RNA-seq analysis. For the time-course in vitro experiment, the two subpopulations were grown for 43 days and pellets were collected at every passage. RT-qPCR was performed on time points every 3-4 passages (n=7). The limiting dilution transplantation experiment *in vivo* was performed using frozen vials of bulk and TM4SF1^{high} cells passage 2 (P2) and propagated in 2D for other two passages (day 9 from sorting). Cells were first checked for subpopulation gene signature expression by qPCR, see Supplementary Figure 6a). **b.** Left: RT-qPCR data reporting the expression of *TM4SF1* in SUM159PT_KRAB cells infected with two sgRNAs targeting *TM4SF1* or a control sgRNA (CTRL) and induced for 72h with doxycycline. Expression levels are normalized to the expression of the housekeeper gene RPLP0. Right: FACS plots representing the percentage of TM4SF1+ cells in control and sgTM4SF1 after 72h of

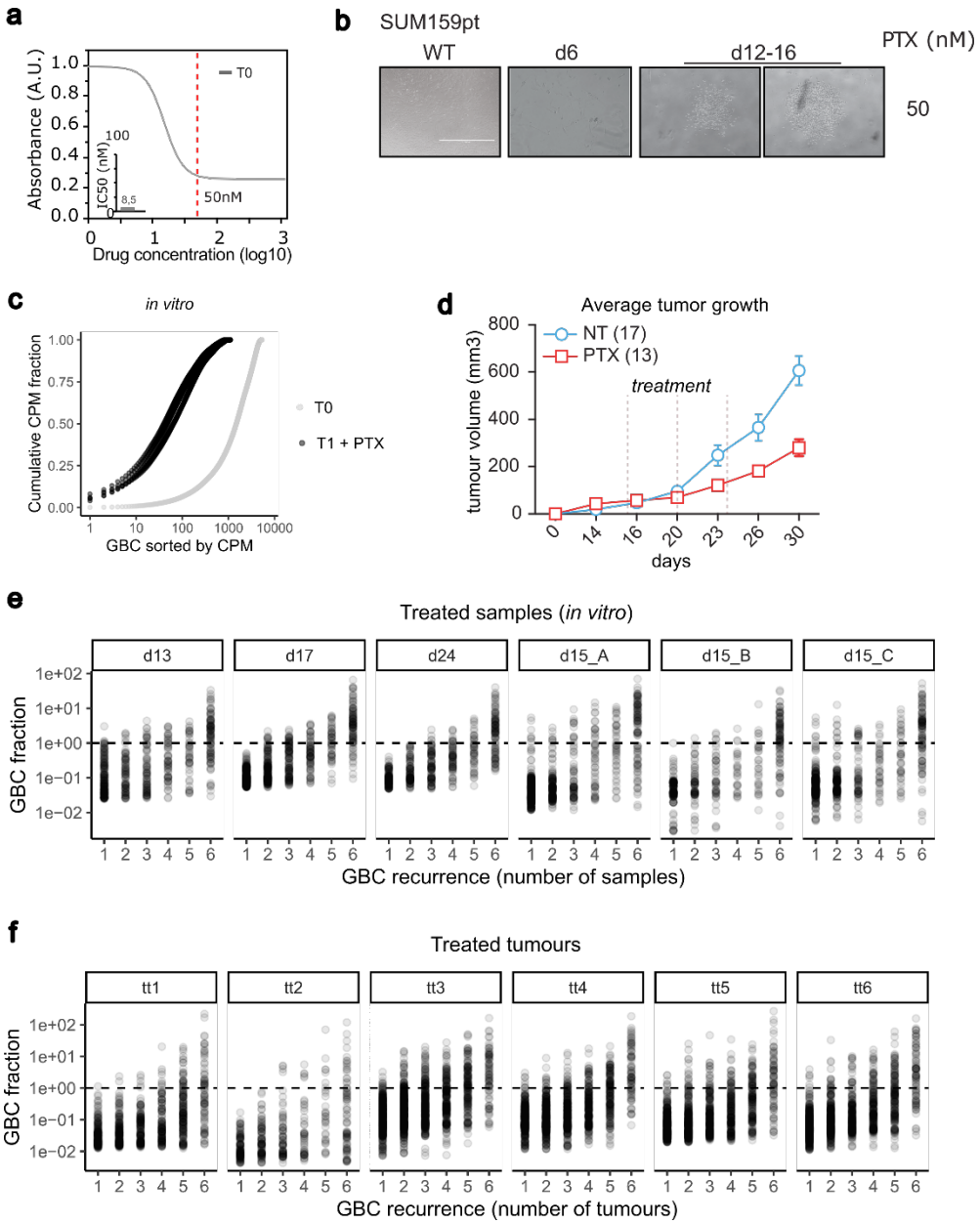
doxycycline treatment and staining with the PC-conjugated TM4SF1 antibody. **c.** Gating strategy for the control experiment in b. **d.** TM4SF1 gating strategy. FACS of live cells were used to separate the TM4SF1high (top 5% and top 10%) subpopulations to establish the gating strategy. Source data are provided as a Source Data file. [a. created with Biorender.com released under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license].



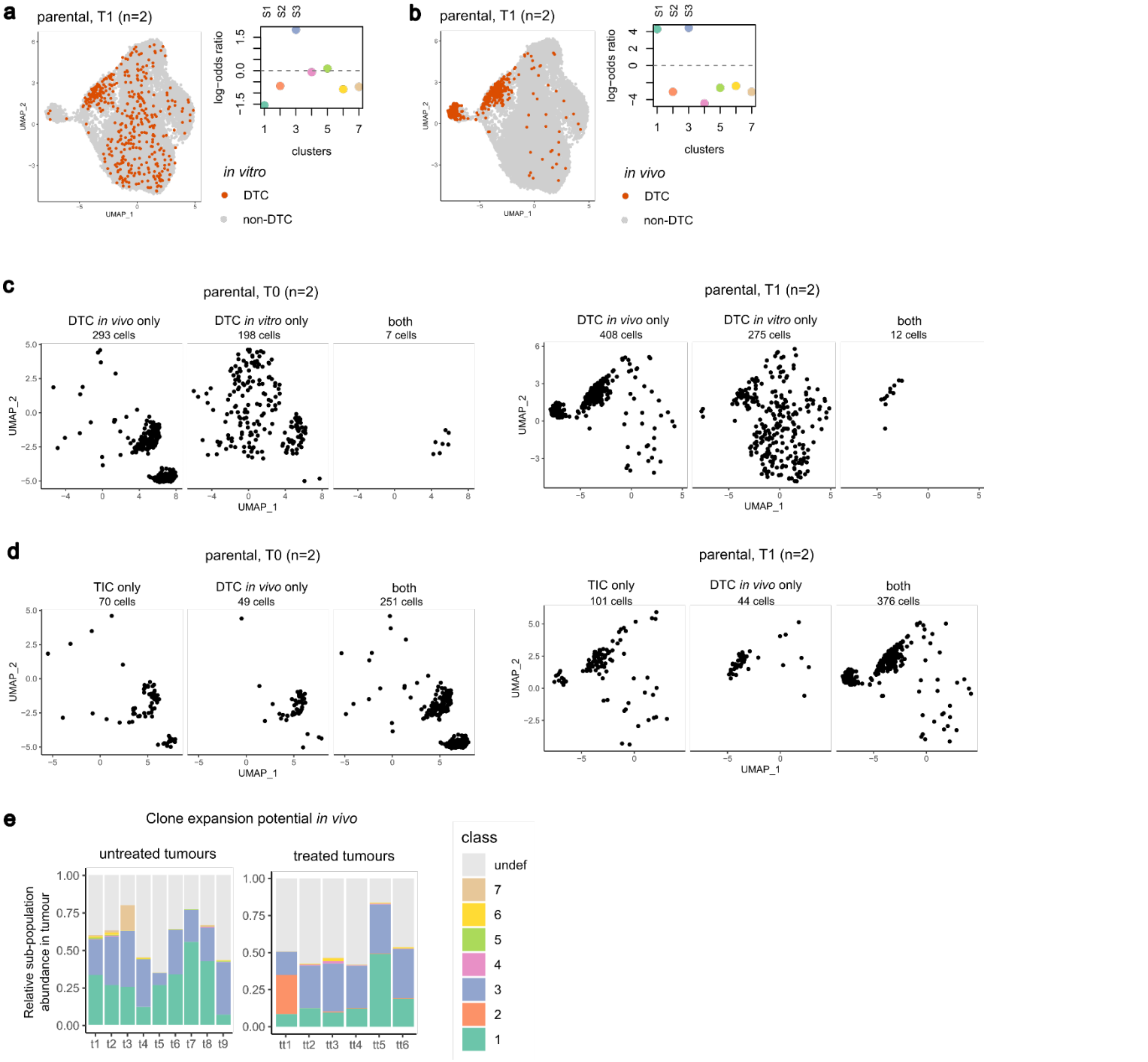
Sorted	Days	Bulk	Days
1A1	32	2A1	49
1A2	35	2A2	49
1A3	35	2A3	ND
1A4	52	2A4	56
1A5	42	2A5	46
1A6	46	2A6	ND
1A7	32	2A7	63

Supplementary Figure 6. Characterization of TM4SF1^{high} cells. a. RT-qPCR data of 7 genes from the S1 gene signature and FEZ1 from the S3 signature respectively in bulk and TM4SF1^{high} cells at time of sorting (day 0) and before in vivo injection (day 9). **b.** RT-qPCR data of the genes shown in a. at top 5% and top 10% TM4SF1 sorting gates, respectively. **c.** Cell plot showing the expression change (RT-qPCR) of selected genes in the S1 signature in a time-course experiment, where TM4SF1^{high} and bulk cells were let grow unperturbed in 2D for several passages. Entries represent the $\log_2$(FC) of each gene at every time point normalized over the expression at the first passage (day 2) in bulk cells. **d.** Comparative functional analysis (by Ingenuity Pathway Analysis, IPA) performed on the 195 upregulated genes shown in Figure 2g, indicating the enriched biological functions (left) and upstream transcription factors (right). Entries satisfying |Z-Score| > 2 and $p < 0.001$ are shown. **e.** Growth dynamics of each individual primary tumour derived from inoculation into mammary fat pads of recipient NSG (NOD/SCID/IL2Rγ$_c^{-/-}$) mice. 100 TM4SF1^{high} cells (at d9) and 100 bulk cells were used, n=7 mice/group. Growth curve after tumours arise until time of euthanisation. **f.** Average weight of mice in response to inoculation with 100 TM4SF1^{high} versus 100 bulk cells in mammary fat pads. Mice weight measurements represent the mean ± SEM, n = 7 mice/group. **g.** Plot showing the latency in days of tumour development for each mouse inoculated with either 100 TM4SF1^{high} or 100 bulk cells. Data

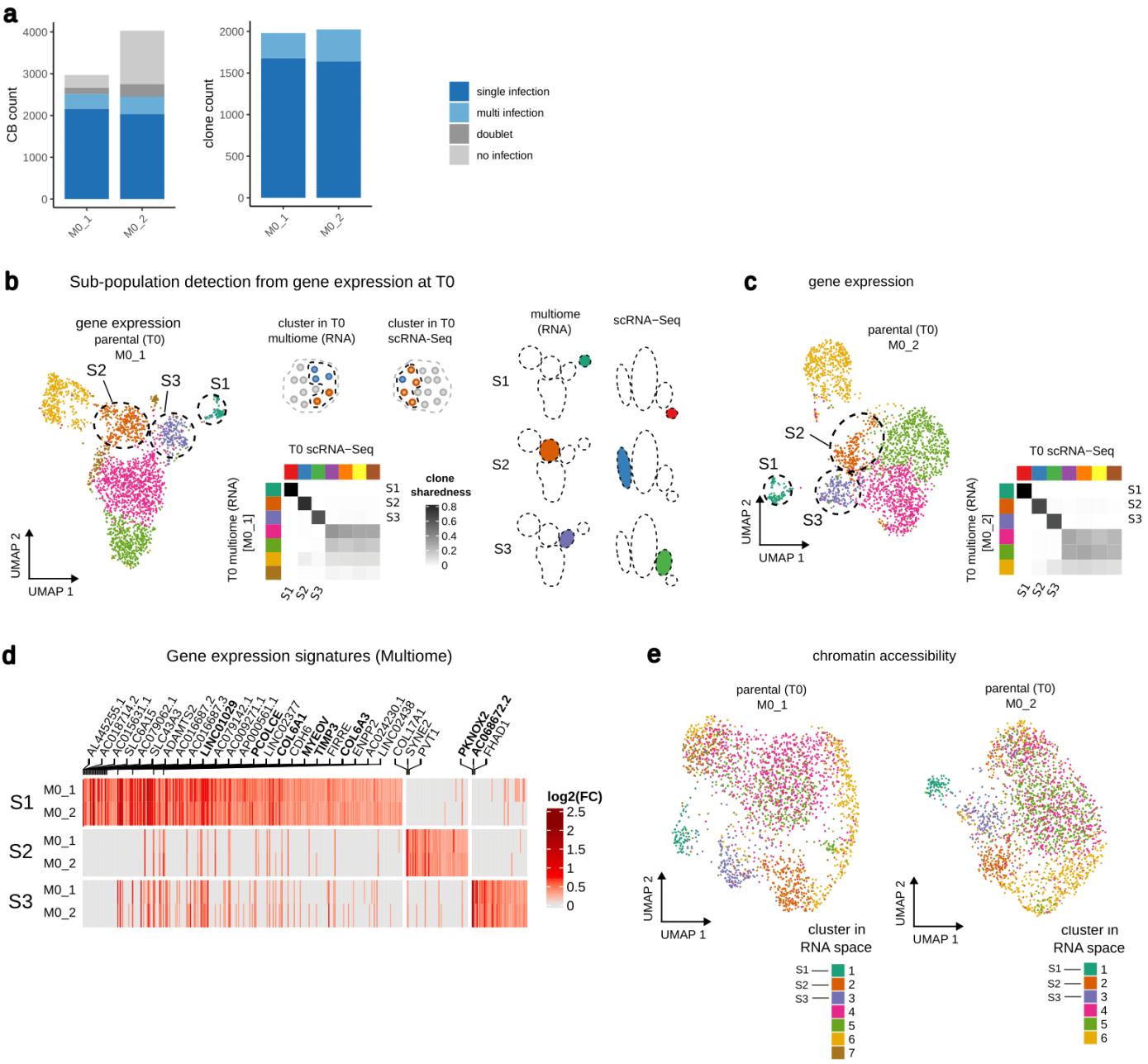
are as in e. *p = 0.0111; TM4SF1^{high} versus bulk cells, two-sided unpaired t-test. Right: latency (days) for each mouse in the two groups, TM4SF1^{high} and bulk. Source data are provided as a Source Data file.



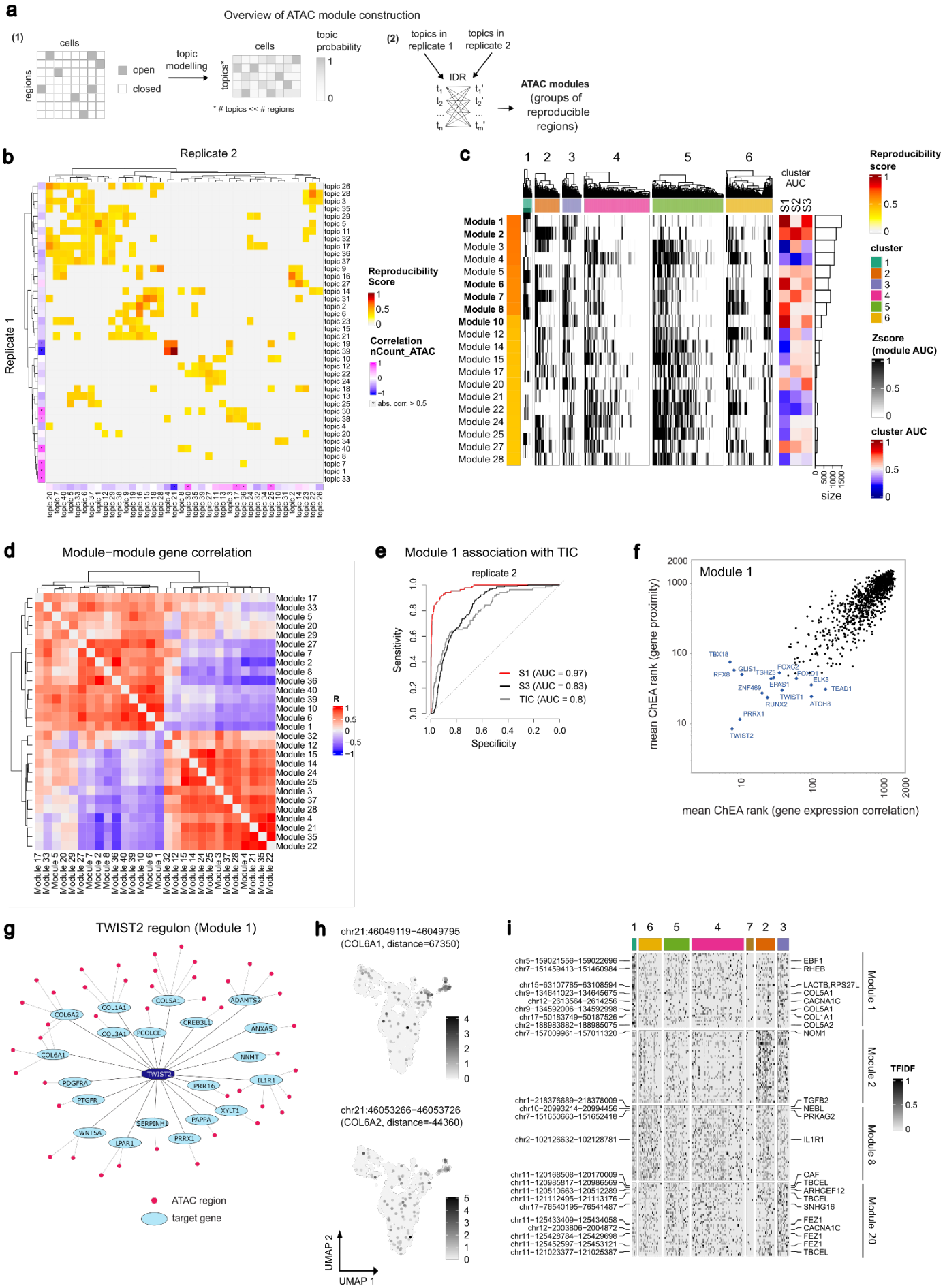
Supplementary Figure 7. Drug tolerant clones are recurrent in tumours. a. Dose response curve of SUM159PT treated with paclitaxel (PTX). The curve was estimated according to the Logistic 4P fit model (JMP software). The value of IC50 is reported in the insert. **b.** Representative bright field images of persister colonies that are typically observed 12 days after paclitaxel treatment of SUM159PT cells. Scale bar 1000 μM. **C.** Clone selection upon paclitaxel treatment (50nM) in the independent infection experiment. Comparison of cumulative clone distribution in parental and treated *in vitro* samples. **d.** Growth dynamics of SUM159PT-derived tumours treated intraperitoneally every 5 days with either PTX (10 mg/kg in PBS, 17 mice) or vehicle (PBS, NT, 13 mice). Each point of the growth curves represents the tumour volume expressed as the mean value ± SEM. Data are from two independent cohorts of mice; at day26 data were collected only in exp2 (NT, 10 mice; PTX, 4 mice). **e.** Relationship between clone abundance and recurrence in 6 paclitaxel-treated sample, late time points (day ≥ 13). Each graph refers to a sample and each dot is a clone (GBC); clones are grouped by the number of times they are observed across the six samples. If a clone is such that $x = k$ in sample s, this means that it is detected in exactly k samples (including s). The y axis is the relative clone abundance in each sample, expressed as the frequency over the total clone count in the sample. **f.** Same as e. for the six paclitaxel-treated tumours (see also Figure 2c legend). Source data are provided as a Source Data file.



Supplementary Figure 8. Drug tolerant clones are associated with the S3 subpopulation. a. Mapping of the drug tolerant clones *in vitro* at parental state (T1). Left: UMAP representation of T1 cells on gene expression space; the 287 cells classified as DTC *in vitro* are coloured in orange. Right: the *x* axis is the cluster identifier, and the *y* axis is the log-odds-ratio obtained from the contingency table comparing cluster assignment and DTC labelling *in vitro* across cells at T1. **b.** Mapping of the drug tolerant clones *in vivo* at parental state (T1). Left: UMAP representation of T1 cells on gene expression space; the 420 cells classified as DTC *in vivo* are coloured in orange. Right: the *x* axis is the cluster identifier, and the *y* axis is the log-odds-ratio obtained from the contingency table comparing cluster assignment and DTC labelling *in vivo* across cells at T1. **c.** UMAP representation of T0 (left) or T1 cells (right), split according to whether cells are classified as DTCs only *in vivo*, only *in vitro*, or in both assays. **d.** UMAP representation of T0 (left) or T1 cells (right), split according to whether cells are classified only as TICs, only as DTCs *in vivo*, or in both assays. **e.** Association between parental state (T1) and clone expansion *in vivo*, without and with treatment. Top: bar plot showing the relative abundance of T0 clusters in every untreated (left) or treated tumour (right), respectively (unassigned clusters are shown in grey). Source data are provided as a Source Data file.

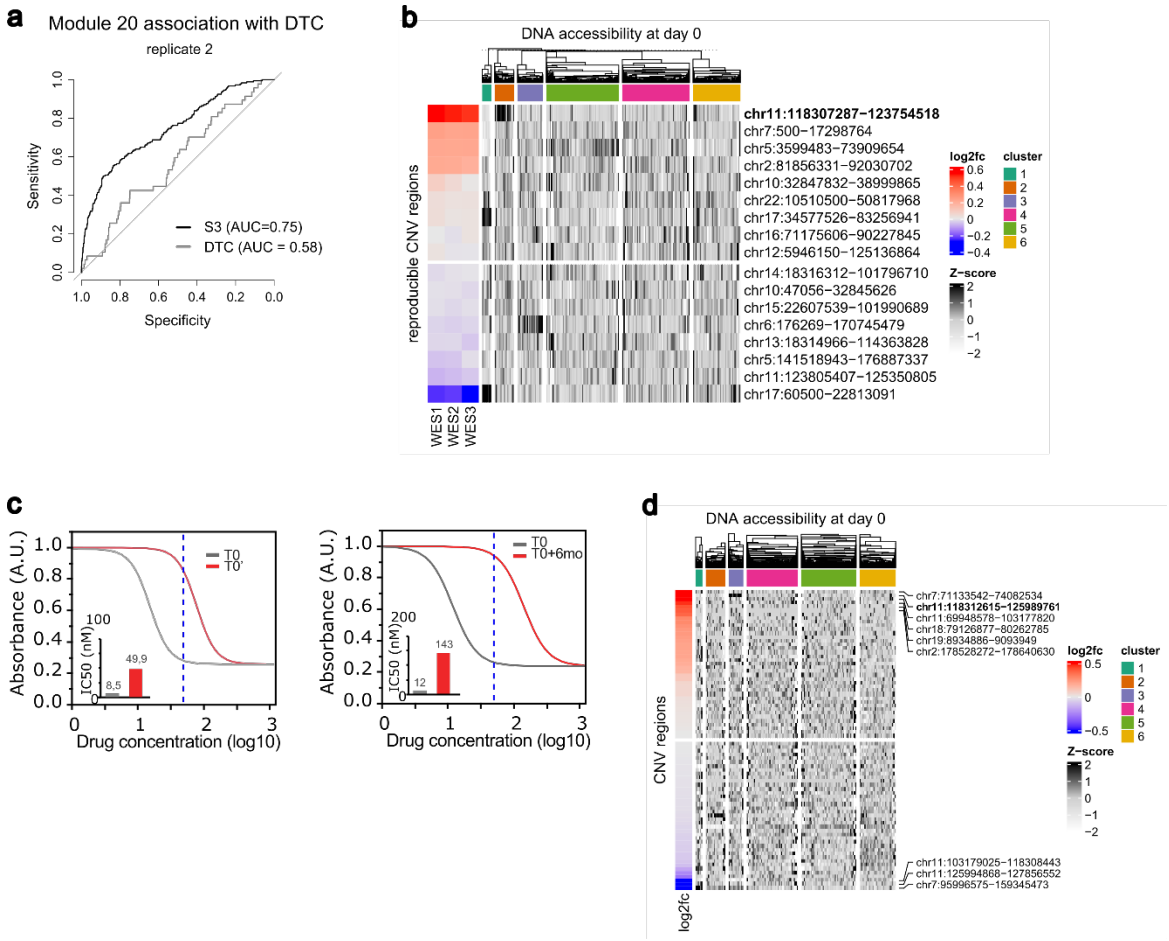


Supplementary Figure 9. Single-cell multiomic analysis of TNBC subpopulations. a. Clone calling for Multiome replicates at T0 (M0_1 and M0_2). Left: cellular barcode (CB) classification into single infection, multi-infection, doublet, and no infection; CB count (left) and clone count (right) are shown (see also Supplementary Figure 1b). **b.** Clone sharedness score between M0_1 and T0. Left: UMAP representation of M0_1 nuclei in gene expression space, coloured by gene expression clusters (2446 nuclei in total). Centre: heatmap where rows are gene expression clusters in M0_1, columns are clusters in T0, and entries are clone sharedness score values for each cluster pair. Rows and columns are sorted according to the pairs with the highest score. The mapping to the subpopulations S1, S2, S3 is indicated for the corresponding clusters in M0_1. Right: schema showing the location of the subpopulations on UMAP. **c.** Clone sharedness score between M0_2 and T0. Left: UMAP representation of M0_2 nuclei in gene expression space, coloured by gene expression clusters (2377 nuclei in total). Right: heatmap where rows are gene expression clusters in M0_2, columns are clusters in T0, and entries are clone sharedness score values for each cluster pair. Rows and columns are sorted according to the pairs with the highest score. The mapping to the subpopulations S1, S2, S3 is indicated for the corresponding clusters in M0_2. **d.** Subpopulation gene signatures in gene expression space (Multiome). Heatmap rows are subpopulations split by sample, columns are genes, and entries are $\log_2$(FC) values between a subpopulation and its complement within the same sample. All differentially expressed genes (DEGs) in at least one subpopulation are reported (see Methods), the top 20 (S1) or top 5 (S2, S3) are labelled with the corresponding gene symbol, and the ones common to scRNA-Seq signatures are highlighted in bold (see also Figure 1g). **e.** UMAP representation on ATAC space, coloured by gene expression clusters for the 2446 nuclei of M0_1 (left) and the 2377 nuclei for M0_2 (right).

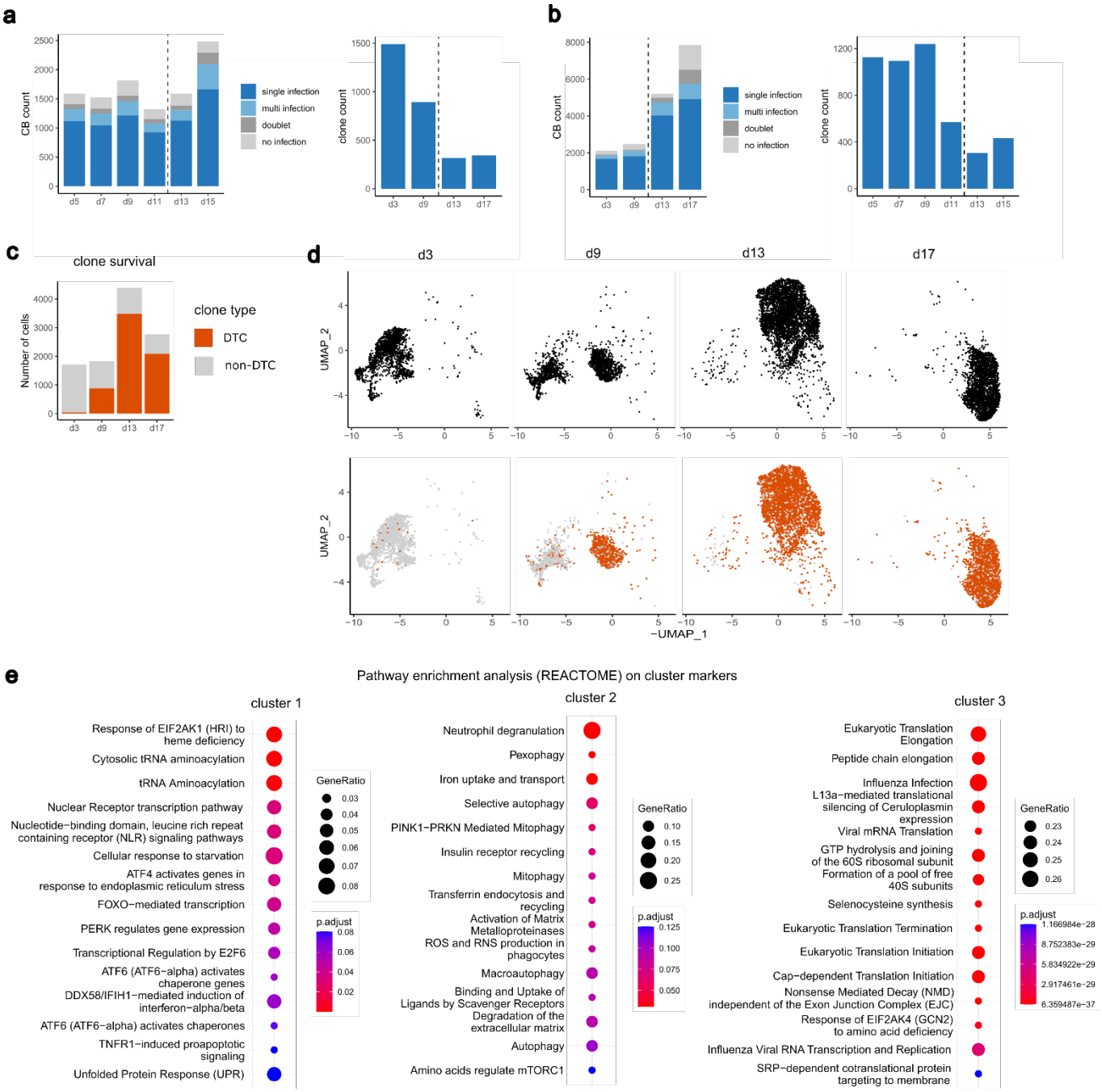


Supplementary Figure 10. Identification and characterization of DNA accessibility modules. a. Procedure for ATAC module extraction. **b.** Comparison of the output of topic modelling on the ATAC regions of the two Multiome

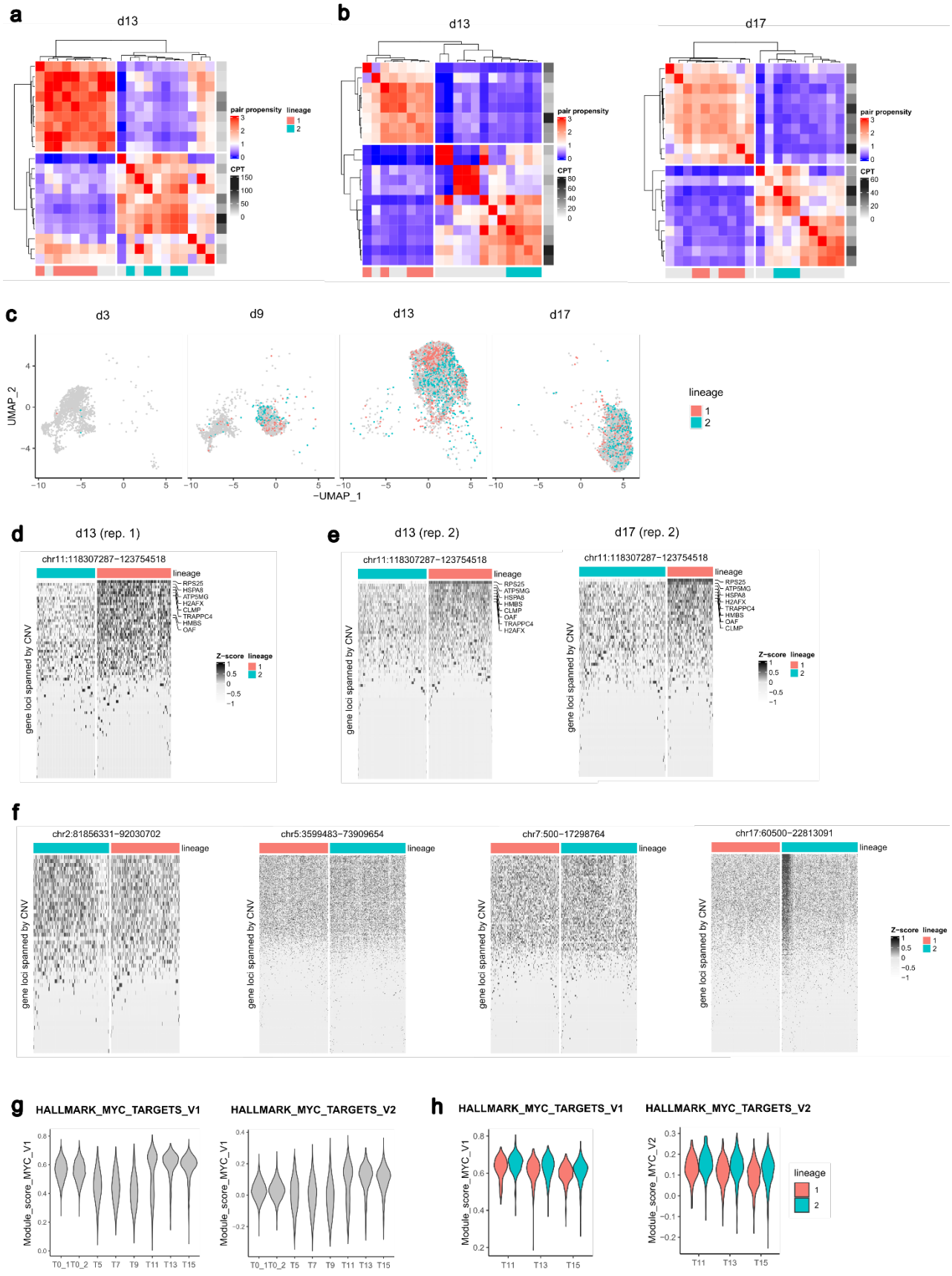
samples. Entries are reproducibility scores (see Methods) and rows and columns are annotated with Pearson's R correlation between (topic,cell) probabilities and ATAC fragment counts (|R| > 0.5 are marked with *). **c.** Comparison between ATAC modules and gene expression subpopulations in single nuclei (M0_2). Heatmap rows are the top 20 scoring modules (≥ 20 regions), columns are nuclei, and entries are module AUC scores representing the overall accessibility of a module in a cell. The association (AUC) of a module with subpopulations (S1, S2, and S3) is shown on the right (blue: negative association; red: positive association); high associations (AUC > 0.75) with any subpopulation are reported in bold. Columns are clustered on Euclidean distances using complete method from hclust package. **D.** Pearson's R correlation between gene scores across ATAC modules (top 40 reproducible modules, ≥ 20 regions). Gene score is the Spearman's rho correlation between gene expression and module AUC. **e.** Module 1 AUC as a predictor of S1 (red), S3 (black), or TICs (grey) on M0_2. **f.** Transcription factor enrichment analysis (average TF rank, ChEA3) on genes positively correlated with Module 1 AUC (*x* axis) or whose locus is located ≤ 100 kbp away from any region in Module 1 (*y* axis). The top 10 ranked genes for either gene set are labelled. **g.** "TWIST2 regulon" including the genes whose expression is positively correlated with Module 1 AUC and that i) are predicted as TWIST2 targets by ChEA3 and ii) lie at <100 kbp from any Module 1 region. **h.** UMAP plot showing the term frequency-inverse document frequency (TFIDF) for the two selected target regions of Module 1 (M0_1). **i.** Top 40 reproducible regions for 4 modules, where rows are regions, columns are nuclei, and entries are TFIDF scores. Columns are clustered on Euclidean distances using complete method from hclust package. Regions located at ≤ 50 kbp from any gene in subpopulation gene signatures (see Figure 1f) are labelled.



Supplementary Figure 11. The chr11 amplification is associated with a drug tolerant phenotype. a. Association between Module 20 and S3 subpopulation (M0_2). Top: ROC curve showing the power of Module 20 AUC score to predict (i) the membership to the DTC in vitro pool (grey line) or (ii) the membership to the S3 subpopulation (black line). **b.** Copy-number variants (CNVs) pre- and post-treatment (drug tolerance assay, M0_2; see Figure 6d). Heatmap showing the association between ATAC-Seq signal from Multiome data and copy-number variants inferred by WES. Rows are consensus CNV obtained from the analysis of paclitaxel-treated samples (day 15; n=3 replicates), as in the drug-tolerance experimental design *in vitro* from Figure 3a; columns are nuclei in M0_2; entries are cumulative ATAC counts in each CNV locus and in each nucleus. Rows correspond to consensus CNVs across three replicates. The coverage $\log_2$(FC) between each treated sample and the untreated reference (baseline SUM159PT cells) is reported on the left; the chromosome location is shown on the right. Columns are grouped by gene expression cluster; hierarchical clustering is performed with complete method from hclust on Euclidean distances. **c.** Dose response curve of different SUM159PT populations (T0, T0' and T0+6months, as defined in Figure 6d-e) treated with paclitaxel. The curve) was estimated according to the Logistic 4P fit model (JMP software). IC50 values are reported in the insert. **d.** CNVs pre- and post-treatment (drug resistance assay, M0_2; see Figure 6e). The heatmap shows the association between ATAC-Seq signal from Multiome data and copy-number variants predicted by WES (see also a.). Rows are CNVs obtained from the analysis of paclitaxel-treated samples (n=1; see Figure 6e). Source data are provided as a Source Data file.



Supplementary Figure 12. Time-course single-cell analysis of paclitaxel response in TNBC cells. a. Clone calling for time-course drug tolerance assay (replicate 1). Left: cellular barcode classification into single infection, multi-infection, doublet, and no infection; CB count (left) and clone count (right) are shown (see also Supplementary Figure 1b). **b.** Clone calling for time-course drug tolerance assay (replicate 2). See a. **c.** Bar plot showing the distribution of DTC *in vitro* on treated samples (replicate 2). **d.** Drug-tolerant clone selection across time (replicate 2). Cells from treated samples are mapped to a common gene expression UMAP space (10698 cells in total), split by sample, coloured depending on whether they are drug tolerant (in orange, as defined in c.) or not (in grey). **e.** Pathway enrichment analysis (REACTOME) of significantly up-regulated genes in clusters of drug-treated samples (as in Figure 7c-d). The top 15 significantly enriched terms (q-value < 0.1) are reported and sorted by non-increasing q-value. The size of the circles is proportional to the fraction of genes found in each pathway for each subpopulation signature.



Supplementary Figure 13. Characterisation of drug tolerant lineages in TNBC cells. a. Pair propensity value of top expanded clone pairs at day 13 (exp1; clones *i* with p_{ii} < 1 are not shown); rows and columns are distinct clones, rows are annotated by clone abundance (in CPT, count per thousand cells), and columns are annotated by lineage. **b.** Pair propensity value for top expanded clone pairs at day 13 and 17 (exp2). See also a. **c.** UMAP representation of cells at day 3, 9, 13 and 17 (exp2) and coloured according to lineage assignment (lineage 1 in pink, lineage 2 in cyan, and unassigned clones in grey). **d.** Heatmap where rows are gene loci spanned by the highest $\log_2$(FC) consensus amplification in drug treated samples, which is located on chromosome 11 (see Figure 6b), columns are

cells at day 13 assigned to lineages (exp2), and entries are log-normalised, scaled UMI counts. Rows are sorted by non-increasing average expression. Columns are split by lineage and clustered with complete method on Euclidean distances. **e.** Heatmap relative to chromosome 11 amplification, where columns are cells at day 17 (replicate 2). See also d. **f.** Heatmaps relative to all the other consensus CNVs such that average $|\log_2(FC)| > 0.1$, for day 15 (replicate 1). See also d. **g.** Quantification of MYC activity. Distribution of MYC activity scores across all time points, before (T0, see Figure 1b) and after treatment (day ≥ 5, see Figure 7a); MYC activity is computed using the Hallmark signatures of MYC Targets (V1 and V2) from The Molecular Signatures Database (MSigDB) with Seurat ModuleScore function. **h.** Distribution of MYC activity score, as in f., at day ≥ 11 and split by lineage.

2. Supplementary References

1. Gavish A*, et al.* Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature* **618**, 598-606 (2023).

2. Pal B*, et al.* A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *EMBO J* **40**, e107333 (2021).

3. Simeonov KP*, et al.* Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states. *Cancer Cell* **39**, 1150-1162 e1159 (2021).

4. LaFave LM*, et al.* Epigenomic State Transitions Characterize Tumor Progression in Mouse Lung Adenocarcinoma. *Cancer Cell* **38**, 212-228 e213 (2020).