Article

Epigenetic heterogeneity in breast tumours informs diagnosis and prognosis

https://doi.org/10.1038/s41586-026-11230-y

Received: 20 May 2025

Accepted: 30 September 2026

Published online: 11 November 2026

Check for updates

Noa Furth[1,23], Christina Riemenschneider[1,23], Yishai Arieli[2,23], Matan Avni[1], Michael-Yakov Povolotski[1], Hadas Yung[1], Shiri Karagach[3], Ofir Atrakchi[3], Nir Erez[1], Tomer-Meir Salame[4], Roberta Noberini[5], Adeline Durand[6], Grégoire Jouault[6], Bareket Dassa[7], Mirie Zerbib[8], Roni Oren[8], Ofra Golani[9], Barak Bar-Zakai[10], Yael Mavor[10], Orli Friedman-Eldar[11], Hadas Kadar-Sfarad[11], Shani Gilron[12], Osnat Givon-Madhala[11], Naama Ronel[13], Faisal Abu-Rmilah[13], Itai Moshe[13], Shahar Ish-Shalom[13], Ady Yosepovich[13], Miriam Kaplan[14], Michal Bakalenik[14], Maya Dadiani[14], Rakel Tryggvadottir[15], Gabriel Shatkin[15], Mariam Hakobyan[16], Alexander Kovacsovics[16], Sebastian D. Mackowiak[16], Amanda Arnold[17], Sarah McLaughlin[18], Ekaterina Andreishcheva[19], Abhijeet Shinde[19], Jiaren Zhang[19], Daniel Jones[19], Einav Nili Gal-Yam[14], Tamar Geiger[3], Tiziana Bonaldi[5,20], Andrew P. Feinberg[15,21,22], Alexander Meissner[16], Celine Vallot[6], Guy Ron[2] ✉ & Efrat Shema[1,16] ✉

Breast cancer, the most diagnosed cancer among women, is stratified by gene expression patterns into luminal and basal-like subtypes, and the latter is associated with worse prognosis[1–3]. Recent advances in single-cell technologies have underscored the presence of intratumour heterogeneity, which poses substantial challenges in cancer treatment and achieving long-term disease-free survival[4–7]. Here we extensively profile epigenetic heterogeneity in a cohort of treatment-naive human luminal tumours, which were obtained directly after surgical resection. By applying a single-cell epigenetic approach based on cytometry by time-of-flight, we identify a recurrent subpopulation of basal-like cells in luminal tumours with a conserved epigenetic signature across patients. These basal-like cells, also identified by deconvolution of whole-genome DNA methylation patterns, are characterized by high expression of aggressive oncogenic traits and downregulation of hormone receptor signalling. Direct comparison of single-cell proteomic and genomic profiling of monomethylation and trimethylation of histone 3 on lysine 4 revealed that these marks are key features associated with loss of luminal identity. Finally, we highlight the potential of concomitant single-molecule profiling of histone modifications on cell-free nucleosomes in blood as an epigenetic liquid biopsy approach for early diagnosis of breast cancer. Moreover, the epigenetic basal signature measured in blood is linked to prognosis, which enables the identification of patients at increased risk for disease relapse.

Gene expression profiles classify breast cancers into luminal A or luminal B (hormone-receptor-positive (HR^+)), HER2-positive and basal-like subtypes. Among these, basal-like subtypes largely overlap with triple-negative breast cancer (TNBC)[2,3]. These classifications guide treatment management and prognosis. Luminal HR^+ breast cancer generally responds to endocrine therapy with favourable outcomes, yet many patients relapse[8]. Single-cell and spatial studies have revealed substantial intratumoural heterogeneity among subtypes, which potentially drive treatment resistance and recurrence[4–7,9].

Chromatin organization, mediated by DNA and histone post-translational modifications (PTMs), facilitates the acquisition of epigenetically distinct cellular states in genetically identical cells[10–12]. Although single-cell RNA sequencing (scRNA-seq) and single-cell chromatin-accessibility sequencing (scATAC–seq) delineate intratumour

[1]Department of Immunology and Regenerative Biology, Weizmann Institute of Science, Rehovot, Israel. [2]Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem, Israel. [3]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [4]Mass Cytometry Unit, Department of Life Sciences Core Facilities, Weizmann Institute of Science, Rehovot, Israel. [5]Department of Experimental Oncology, European Institute of Oncology IRCCS, Milan, Italy. [6]UMR3244, Translational Department, Institut Curie, PSL University, Paris, France. [7]Bioinformatics Unit, Department of Life Sciences Core Facilities, Weizmann Institute of Science, Rehovot, Israel. [8]Department of Veterinary Resources, Weizmann Institute of Science, Rehovot, Israel. [9]Department of Life Sciences Core Facilities, Weizmann Institute of Science, Rehovot, Israel. [10]Department of Surgery A, Kaplan Medical Center, Rehovot, Israel. [11]The Center for Breast Health, Kaplan Medical Center, Rehovot, Israel. [12]Plastic and Reconstructive Surgery, Kaplan Medical Center, Rehovot, Israel. [13]Department of Pathology, Kaplan Medical Center, Rehovot, Israel. [14]Breast Oncology Institute, Sheba Medical Center, Tel Aviv, Israel. [15]Department of Biomedical Engineering, Johns Hopkins Whiting School of Engineering, Baltimore, MD, USA. [16]Max Planck Institute for Molecular Genetics, Berlin, Germany. [17]Cancer Clinical Studies Unit, Mayo Clinic, Jacksonville, FL, USA. [18]Department of Surgery, Mayo Clinic, Jacksonville, FL, USA. [19]Epivision Genetics, Boston, MA, USA. [20]Department of Oncology and Haemato-Oncology, University of Milan, Milan, Italy. [21]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [22]Center for Epigenetics, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [23]These authors contributed equally: Noa Furth, Christina Riemenschneider, Yishai Arieli. ✉e-mail: guy.ron2@mail.huji.ac.il; efrat.shema@weizmann.ac.il

heterogeneity[13–16], the contribution of chromatin patterns remains unclear[17]. Single-cell profiling of histone marks, such as trimethylation of histone 3 on lysine 27 (H3K27me3), implicates chromatin states in breast tumour heterogeneity and treatment resistance[18,19]. However, these methods measure a limited number of chromatin marks, which hinders analyses of complex patterns and interactions. We developed a cytometry by time-of-flight (CyTOF) method to systematically profile multiple histone PTMs and decode population-specific chromatin patterns to examine heterogeneity among cancer cells[20–23].

Tumour epigenetic states are also reflected in blood. Cell-free, plasma-circulating nucleosomes, released mainly by dying normal and cancer cells, retain tissue-of-origin epigenetic patterns[24–27]. Our single-molecule imaging platform, epigenetics of plasma isolated nucleosomes (EPINUC), captures multiple epigenetic features on plasma nucleosomes[28]. This method detects colorectal cancer, including low-burden early-stage disease[28], and enables minimally invasive molecular profiling and treatment monitoring of paediatric brain tumours[29].

Here we present a holistic epigenetic analysis of primary treatment-naive breast tumours and blood cell-free nucleosomes. Linking recurrent intratumoural epigenetic patterns to circulating nucleosomes connects tumour epigenetic heterogeneity with its potential clinical implications. This work reveals epigenetic alterations as key mediators of intratumour heterogeneity and potential biomarkers for early diagnosis and prognosis.

Single-cell epigenetics of breast tumours

To assess heterogeneity in breast tumour histone modifications, we performed epigenetic focused CyTOF on fresh, surgically resected tumours from treatment-naive women and quantified histone modifications at single-cell resolution (Fig. 1a and Supplementary Table 1). The cohort included mainly luminal HR^+ tumours and only one basal triple-negative (TN, HR^-) tumour ($n = 12$; Supplementary Table 2 and Supplementary Fig. 1a,b), as standard neoadjuvant therapy excluded most TN cases from treatment-naive eligibility. Following batch correction and normalization, uniform manifold approximation and projection (UMAP) analysis of all 12 tumours revealed that tumour cells largely occupied patient-specific regions, whereas stromal cells clustered together across patients. This result aligns with previous studies that measured single-cell genomic differences[13,14] (Fig. 1b, Methods and Supplementary Fig. 2a). Partial overlap among luminal tumour cells from different patients probably reflects shared lineage-specific epigenetic programs. As expected, stromal cells expressed vimentin and αSMA, whereas tumour cells expressed the epithelial keratins, EpCAM and E-cadherin (Fig. 1b,c and Supplementary Fig. 2b,c). Notably, the basal TN tumour differed from HR^+ tumours, showing low expression of the luminal marker GATA3 and high expression of the basal marker cytokeratin-5 (K5; also known as KRT5) and mesenchymal vimentin markers (Fig. 1b,c and Supplementary Fig. 2d). Stromal and tumour chromatin patterns also differed, a result that underscores their role in dictating cellular identity (Fig. 1d and Supplementary Fig. 2e). CyTOF analysis of MCF7 (luminal) and four basal-like TN cell lines validated that GATA3 and K5 expression patterns can differentiate luminal and basal-like identities. Moreover, SHAP analysis[30,31] identified these two markers as the strongest discriminators of tumour type (Fig. 1e,f). Using multiple cell-identity markers, we derived a PermCell luminal signature (Methods) that accurately distinguished the identities of the cell lines (Fig. 1g).

We next examined intratumour heterogeneity, focusing on the cancer cells of each luminal breast tumour separately. UMAP clustering and density-based spatial clustering of applications with noise (DBSCAN) across all features revealed three subpopulations in each tumour, which we termed cycling, luminal and basal-like (Fig. 1h,i and Extended Data Fig. 1a,b). These subpopulations were also conserved in two different luminal patient-derived xenograft (PDX) models. This finding validates the observed heterogeneity as an inherent feature of luminal breast tumours and excludes contamination of normal cells (Fig. 1j and Extended Data Fig. 1c). MCF7 cells showed similar luminal and basal-like heterogeneity, although most cells were cycling, as expected for a cell line (Extended Data Fig. 1d). In tumours, only a luminal subset cycled and expressed high levels of Ki-67, a result consistent with previous immunohistochemistry (IHC) and scRNA-seq studies[14,32]. CyTOF Ki-67 levels correlated with pathology estimates (Fig. 1k). Cycling cells also showed high levels of multiple chromatin modifications, both in raw measurements and following normalization to core histones (Fig. 1i, Extended Data Fig. 1a–c and Supplementary Fig. 3). We validated this result by repeating CyTOF analyses of two tumours (14 and 18), adding the established cell cycle marker phospho-Rb (pRb). pRb-positive cells showed an increase in core histones, as expected from cycling cells, and a larger increase in histone PTMs (Supplementary Fig. 3). Accordingly, the epigenetically inferred cycling fraction strongly correlated with Ki-67 levels (Fig. 1l).

Most cancer cells strongly expressed GATA3 and scored highly for the luminal signature, consistent with clinical classification (Fig. 1h, Extended Data Fig. 1a,b and Supplementary Fig. 4a). Moreover, in each tumour, the PDX samples and the MCF7 cells, we identified a minor recurrent subpopulation that acquired basal-like identity, with high K5 expression and a reduced luminal signature (Fig. 1h–j, Extended Data Fig. 1 and Supplementary Fig. 4a). This phenotypic diversity is consistent with previous single-cell studies of human and mouse breast tumours[14,21,33,34]. Notably, luminal and basal-like cells also had distinct histone PTM signatures (Fig. 1h–j and Extended Data Fig. 1). Accordingly, UMAP plots based solely on histone modifications revealed separation of the subpopulations, which validated the epigenetic basis of this phenotypic diversity (Supplementary Fig. 4b). In MCF7 cells, epigenetic UMAP and clustering similarly identified several clusters, including a $K5^{hi}$ basal-like population (Fig. 1m). These basal-like cells with reduced luminal signalling appeared closer in UMAP space to the TN cell lines, which provided further support for their basal-like identity (Supplementary Fig. 4c). Independent CyTOF experiments, additional MCF7 parental cell culture and populations derived from independent single-cell clones revealed similar subpopulation patterns, a result that argues against a genetic basis for these effects (Fig. 1n and Extended Data Fig. 2a–c). Moreover, individual PDX tumours derived from the same donor and grown independently in different mice showed similar subpopulation proportions, which provides support for a dynamic and plastic relationship between cell states (Extended Data Fig. 2d). Overall, our data revealed inherent intratumour heterogeneity in luminal tumours, with epigenetic variability between distinct subpopulations.

Epigenetically distinct basal-like cells in tumours

We further explored the characteristics of the cycling, luminal and basal-like subpopulations. Unsupervised clustering grouped corresponding subpopulations across patients, which confirmed their robustness and similarity (Fig. 2a). Across tumours, luminal cells expressed oestrogen receptor (ER), GATA3 and cytokeratin-8 and cytokeratin-18 (K8/18; also known as KRT8 and KRT18, respectively), whereas basal-like cells downregulated these luminal markers and highly expressed K5 (Fig. 2b). As these subpopulations also exhibited distinct and conserved chromatin patterns, we applied a machine-learning algorithm to predict intratumour classification[35]. In each tumour, a model trained on a random subset predicted the remaining cells with 94% accuracy, which confirmed the robust differences between subpopulations (Fig. 2c and Extended Data Fig. 3a). It also assigned 78% of unclustered 'noise' cells with >95% confidence, generating classifications that aligned with the original clustering and UMAP positions (Extended Data Fig. 3b). SHAP analysis identified the features driving each subpopulation. In detail, histone PTMs, particularly the four acetylation marks, were leading predictors of cycling cells despite overlap of individual marks with basal-like cells (Extended Data

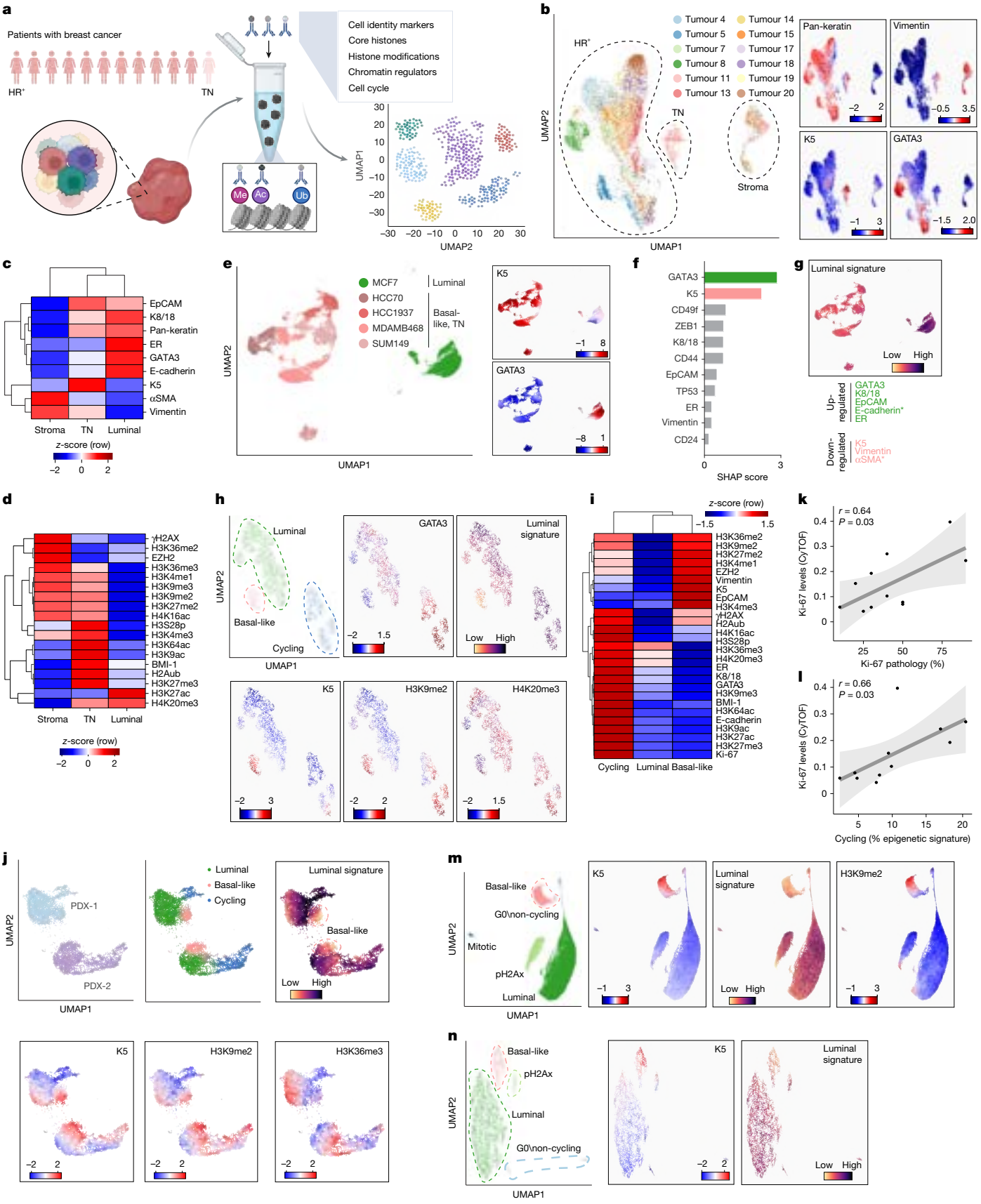


Fig. 1 | See next page for caption.

Fig. 1 | Epigenetic-focused CyTOF analysis of human breast tumours reveals intertumour and intratumour heterogeneity. a, Schematic of the experiment. Tumours were dissociated, depleted of CD45$^+$ immune cells, stained with metal-conjugated antibodies and analysed by CyTOF (Supplementary Table 1). Ac, acetylation; Me, Methylation; Ub, Ubiquitination. **b**, UMAP of the CyTOF dataset (1,157,765 cells, $n = 12$ tumours, mean of 96,480 cells per tumour; Supplementary Table 3). Shown are around 1,000 cells per sample (one replicate per tumour), coloured by sample (left) or marker intensity (right). Dashed outlines indicate HR$^+$, TN and stromal cell regions. **c**,**d**, Heatmaps of mean marker levels in stroma, TN tumour ($n = 1$) and luminal tumours ($n = 11$); rows are standardized and clustered. **e**, UMAP of breast cancer cell lines. Shown are 5,000 cells per line, coloured by line (left) or marker intensity (right). **f**, XGBoost classifier segregating luminal and basal-like cell lines using all cell identity markers (Supplementary Table 1). SHAP scores indicate feature importance. **g**, UMAP as in **e**, coloured by a luminal signature score derived from the indicated markers (bottom). Asterisk indicates markers absent from the cell-line panel. **h**, UMAP of cells from tumour 5 (representative of $n = 11$) following DBSCAN clustering and noise (unclassified cells) removal, coloured by subpopulation, marker intensity and luminal signature. **i**, Heatmap of mean marker levels per subpopulation in **h**, row-standardized as z-scores. Rows and columns are clustered. **j**, UMAP of two luminal PDX models, coloured as in **h**. For PDX-1, six tumours were sampled equally to match PDX-2 cell numbers. **k**,**l**, Mean CyTOF Ki-67 levels versus the percentage of Ki-67$^+$ cells by IHC (**k**) or cycling cells by epigenetic signature (**l**, $n = 11$). Data show linear fit (lines), 95% confidence intervals (CI, grey), Pearson's r and two-sided P values. Tumour 20 showed outlier CyTOF Ki-67 levels and was excluded. **m**,**n**, UMAP and DBSCAN clustering of MCF7 parental cells (**m**) and a single-cell-derived clone (**n**, clone 5, using only epigenetic features, coloured as in **h**.Results were reproduced in seven independent MCF7 CyTOF experiments and three independently derived clones (Extended Data Fig. 2). Schematic in **a** created in BioRender; Furth, N. https://biorender.com/twf1psb (2026).

Fig. 3c,d). Polycomb-associated H3K27me3 and H2Aub also contributed to the features, a finding consistent with previously observed Polycomb deregulation and pro-proliferative activity in breast cancer[36–38].

Luminal and basal-like separation was similarly epigenetic driven, with H4K20me3, H3K36me3 and H3K9me2 marks having the highest SHAP scores (Fig. 2d). Across tumours, basal-like cells consistently had increased H3K9me2 and H3K4 monomethylation and trimethylation, whereas luminal cells had enhanced H4K20me3, H3K36me3, H3K9me3 and H3K27me3 marks (Fig. 2e). These results align with subtype-associated histone PTM signatures[39] and reported roles for H3K4 demethylases, such as JARID1B, in luminal–basal transitions[40]. Both luminal and basal-like cells showed similar pRb distribution, which provide support for the conclusion that these results reflect distinct cell states rather than cell cycle differences (Extended Data Fig. 3e). Similar epigenetic signatures distinguished luminal and basal-like populations in PDX models (Extended Data Fig. 3f). Accordingly, epigenetic-only basal-like PermCell scores accurately captured basal-like cells in tumours, PDXs and MCF7 cells (Fig. 2f and Extended Data Fig. 4). Overall, our computational framework confirms that there is recurrent epigenetic heterogeneity in HR$^+$ breast tumours and highlight chromatin patterns associated with basal-like K5^{hi} cells.

Consistent with the CyTOF results, tumours with higher fractions of epigenetically defined basal-like cells had fewer ER-positive and progesterone receptor (PR)-positive cells, as measured by IHC (Extended Data Fig. 5a,b and Supplementary Table 2). Basal-like cells were also more abundant in worse-prognosis luminal B than luminal A tumours. This result is in agreement with studies showing that these tumours exhibit features shared with the basal-like subtype[3,41–43] (Fig. 2g). Although tumour mutational burden was low across the cohort, it correlated positively with the CyTOF-defined basal-like fraction, which parallels the greater genomic instability typically associated with TNBC[44,45] (Fig. 2h and Supplementary Table 2). TP53 mutant tumours likewise contained more basal-like cells (Extended Data Fig. 5c), a finding reminiscent of single-cell studies of glioblastoma showing that genetic mutations may affect balance between cellular states[46]. Overall, these associations between pathological and molecular characteristics link the epigenetically distinct basal-like subpopulation to a more aggressive disease phenotype.

To characterize these cells functionally, we performed FACS to sort cells from tumour 19 and then low-input mass spectrometry. After excluding stroma through pan-cytokeratin staining, cells were sorted according to K5 levels to enrich for basal-like cells compared with the luminal population. K5^{low} cells highly expressed luminal markers and proteins associated with oestrogen signalling, whereas basal-like K5^{hi} cells expressed proteins involved in oncogenic pathways, such as MYC–mTOR signalling (Fig. 2i–k). Moreover, multiple chromatin regulators differed between these subpopulations, including the H3K4me1 methyltransferase SETD7 and histone demethylase KDM4A, consistent with the histone PTM differences observed by CyTOF[47,48] (Extended Data Fig. 5d). KDM4A, enriched in K5^{hi} cells, targets H3K9me3 and H3K36me3, which potentially accounts for their reduced levels and H3K9me2 accumulation in the basal-like population[49]. Comparisons with a normal mammary single-cell atlas[50] revealed hormone-sensing programs in K5^{low} cells and an alveolar progenitor signature in K5^{hi} cells (Fig. 2l), a result consistent with their enrichment in luminal and basal TN tumours, respectively[50].

To gain a mechanistic understanding of these findings, we sorted MCF7 cells using K5-specific and GATA3-specific antibodies and performed transcriptomic profiling, comparing basal-like cells (K5^{hi}GATA3^{low}) to bulk MCF7 or extreme luminal cells (K5^{low}GATA3^{hi}; Extended Data Fig. 5e,f). Notably, basal-like cells were enriched for epithelial-to-mesenchymal transition (EMT) genes ($P < 0.05$; Extended Data Fig. 5f). We then focused on differentially expressed cell surface markers to enable live sorting, which revealed that CD74 was significantly upregulated in basal-like cells in both MCF7 and tumour 19 proteomic analyses (Fig. 2m and Extended Data Fig. 5g). Live-sorted CD74^{hi} cells exhibited increased migratory capacity in migration and wound-healing assays, which provided further support for the more aggressive phenotypes of basal-like cells (Fig. 2n,o and Extended Data Fig. 5h). Together, these single-cell and bulk proteomics and functional assays identify epigenetically distinct K5^{hi} basal-like cells in luminal tumours associated with aggressive traits such as reduced luminal signalling, elevated oncogenic signalling and increased migratory potential.

DNA methylation reveals intra-tumor heterogeneity

The distinct chromatin and phenotypic states of the basal-like subpopulation led us to explore whether intratumour epigenetic heterogeneity is also apparent in DNA methylation patterns. Thus, we conducted deep whole-genome bisulfite sequencing (WGBS) on three luminal tumours from our cohort. DNA methylation patterns accurately capture breast cancer subtyping, as shown by The Cancer Genome Atlas (TCGA)[51]. We next jointly deconvoluted the TCGA and our WGBS data into functional archetypes (V1–V3), with each tumour positioned in the deconvolution space. Stratifying the data into three archetypes revealed the expected separation of basal (TN) tumours from luminal (HR$^+$), with archetype V3 associated with basal tumours (Extended Data Fig. 6a). Tumours 18, 19 and 23 from our cohort clustered with the luminal tumours, as expected from their bulk clinical subtyping. Correlation of the archetype contribution with RNA expression confirmed the association of V1 and V2 with luminal transcriptional programs, whereas V3 was associated with EMT, a result consistent with previous studies of TN tumours[52–54] (Extended Data Fig. 6b and Supplementary Fig. 5a). To improve resolution, we added a fourth archetype (V4). This analysis revealed another layer of diversity: basal TN tumours were strongly associated with V3, which corresponded with reduced oestrogen signalling and increased

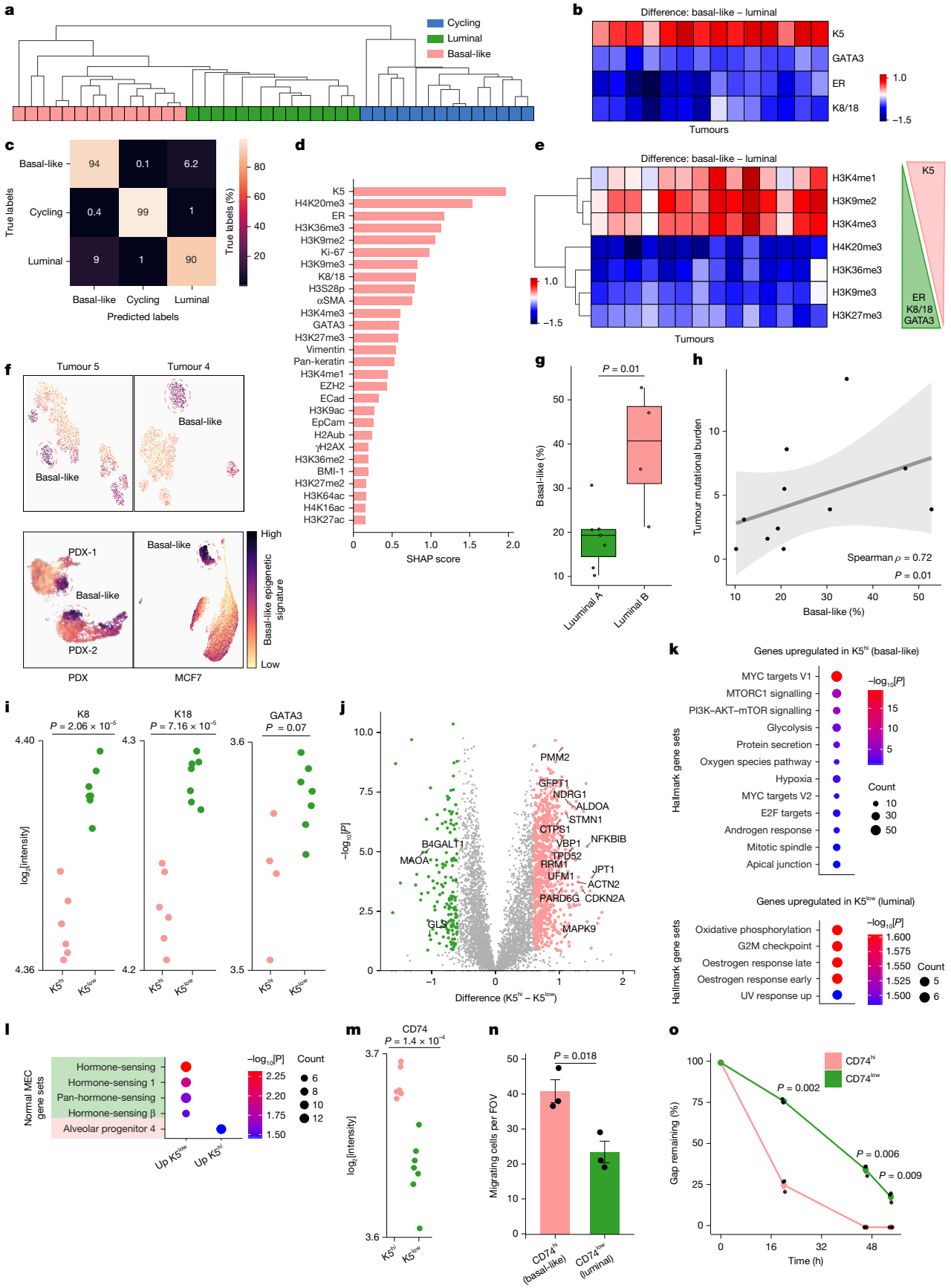


Fig. 2 | See next page for caption.

Fig. 2 | Luminal breast tumours contain basal-like cells associated with distinct epigenetic signature and pro-oncogenic features. a, The three subpopulations from each HR$^+$ tumour ($n = 14$, including 3 anchor repeats; Supplementary Table 3) clustered by CyTOF features. The dendrogram shows recurrent patterns across patients. **b**, Heatmap of mean cell identity marker differences between basal-like and luminal subpopulations ($n = 14$, as in **a**). **c**, Confusion matrix of XGBoost predictions of subpopulations (HR$^+$ tumours, $n = 14$). Rows are normalized per true class, with 94% accuracy. **d**, SHAP scores for feature contribution to basal-like versus luminal classification, ranked in descending order. XGBoost accuracy = 92%. **e**, Heatmap as in **b** of the top discriminating histone PTMs shown in **d**; rows are clustered. **f**, UMAP plots of tumours 4 and 5, luminal PDX models ($n = 2$) and MCF7 cells, coloured by the basal-like epigenetic signature (Extended Data Fig. 4a). The dashed line indicates the basal-like population. **g**, Percentage of basal-like cells in HR$^+$ tumours stratified by luminal A ($n = 7$) or luminal B ($n = 4$) subtype. Boxes: median and 25–75th percentiles; whiskers indicate 1.5× interquartile range (IQR); and points are individual tumours. Two-sided Wilcoxon rank-sum test. **h**, Percentage of basal-like cells versus tumour mutational burden ($n = 11$). Data show the linear fit (line), 95% CI (grey shading), Spearman's ρ and two-sided P value. **i**, Levels of luminal identity markers in K5^{hi} and K5^{low} cells from tumour 19. Dots indicate protein intensity ($\log_2$) in technical repeats of 200 sorted cells (6–7 wells per population). Two-sided Welch's t-test, with no adjustment for multiple comparisons. **j**, Volcano plot of protein expression differences between K5^{hi} and K5^{low} cells, as in **i**. Axes show protein intensity difference and $-\log_{10}[P]$. P values (two-sided Student's t-tests) were corrected by false-discovery rate (FDR). Green and pink depict proteins significantly deregulated (fold change > 1.5, $q < 0.05$). Labelled proteins have fold change >2 and correspond to enriched terms in **k**. **k**,**l**, Functional enrichment of proteins significantly upregulated defined in **j**. **k**, MSigDB hallmark terms. **l**, Mammary epithelial cell (MEC) lineages[50] (Supplementary Table 4a). Terms with nominal $P < 0.05$ are shown (one-sided hypergeometric test). Dot size indicates overlapping proteins, colour $-\log_{10}[P]$. **m**, CD74 protein expression as in **i**. **n**–**o**, MCF7 cells were sorted by CD74 levels and assayed for Transwell migration towards FBS and EGF (**n**) or wound healing (**o**), using $n = 3$ independently recovered sorted populations from the same experiment. For **o**, the gap area was quantified relative to day 0. Data are the mean ± s.e.m.; dots indicate replicates. Two-sided Welch's t-test, Holm-corrected.

oncogenic signalling, whereas luminal tumours were more heterogeneous and spread across the V1 axis (Fig. 3a,b and Supplementary Fig. 5b). Notably, V1 was strongly enriched for EMT, thereby potentially capturing an intermediate-transitional state between luminal and basal-like programs. Moreover, the percentage of CyTOF-measured, vimentin-positive cells, indicative of EMT, in tumours 18 and 19 differed accordingly (Fig. 3c and Supplementary Fig. 5c). Tumour 19, with a higher V1 contribution, also comprised more vimentin-positive cells, as assessed by CyTOF. These results support the use of WGBS deconvolution to capture functionally distinct cellular states in tumours, such as luminal cells undergoing EMT.

To directly link the CyTOF-defined subpopulations with DNA methylation patterns, we sorted K5^{hi} and K5^{low} cells from tumour 5 and tumour 23, followed by enzyme-mediated methylation sequencing (EM-seq). Functional enrichment of differentially methylated promoters revealed hypermethylation of oestrogen-response programs in K5^{hi} cells, whereas genes associated with aggressive and oncogenic features, including KRAS signalling and mesenchymal or basal-state transitions, were hypomethylated (Extended Data Fig. 6c,d). We next derived a CpG island (CGI) methylation signature to distinguish the two states using CGIs that were concordantly differentially methylated across both tumours. In TCGA breast cancer data, this signature scored higher in basal-like tumours, which indicated that the K5^{hi} methylation program resembles basal-like tumours (Extended Data Fig. 6e). K5^{hi} samples also showed higher contribution to the basal-associated and EMT-associated archetypes than K5^{low} samples, in both the three-archetype and four-archetype models (Extended Data Fig. 6f,g). Thus, the CyTOF-defined basal-like population exhibits basal-like and EMT-related methylation programs, which reveals another layer of intratumoural epigenetic heterogeneity.

Basal-like cells have a noisy epigenetic landscape

Having identified distinct DNA methylation programs associated with tumour cell states, we next assessed cell-to-cell variability by clustering read-level methylation patterns across 100-bp genomic bins and calculating empirical entropy as an estimation of cell-type heterogeneity[55]. Genes near high-entropy bins were associated with functional terms related to poor differentiation, tumour aggressiveness and basal compared with luminal identity (Supplementary Fig. 6a). This result suggests that empirical entropy, by capturing diverse DNA methylation states, can be used as a proxy for cell-to-cell heterogeneity and reveal functional differences between subpopulations. We next calculated the mean entropy for each CGI (Extended Data Fig. 7a). As expected, most CGIs were unmethylated across all cells[56], thus showing zero entropy. The variable CGIs (entropy > 0) were functionally enriched for EMT and the related Hedgehog signalling pathways, which indicated that these are differentially regulated in tumours[57] (Extended Data Fig. 7b). This analysis suggests that intratumour variability in methylation patterns may be associated with cell identity transitions.

To further explore intratumour entropy and its relation to cell identity, we leveraged a published multiomic study of luminal breast cancer that included single-cell expression and chromatin accessibility measurements[13]. Focusing on cancer cells, we observed intratumour heterogeneity in luminal and basal-like expression patterns, as previously described[14,54] (Supplementary Fig. 6b). Using the ATAC–seq modality, we calculated the within-peak (WIP) fraction, a metric of epigenetic noise[58], whereby a low-WIP fraction may indicate a destabilized or noisier epigenome, which supports cell plasticity. For each tumour, we correlated the WIP fraction with chromatin accessibility of basal-like and luminal gene signatures. Cells that scored high for EMT and basal-like identity showed lower WIP fraction, a finding reminiscent of the higher DNA methylation entropy associated with these signatures (Fig. 3d and Supplementary Fig. 6c). Single-cell multiome analysis of a freshly resected independent luminal tumour reproduced the basal–luminal continuum and association between basal identity and reduced WIP fraction (Supplementary Fig. 6d,e). To explore whether the increased epigenetic noise is associated with higher transcriptional heterogeneity, we divided cells into quartiles according to a published basal–luminal index[59] and basal-like and luminal gene sets (Methods). Indeed, across all tumours, cells with a high basal identity showed higher transcriptional heterogeneity (Extended Data Fig. 7c). In line with these observations, the CyTOF-defined basal-like subpopulation also showed higher epigenetic heterogeneity (Extended Data Fig. 7d). Moreover, read-level methylation entropy was higher in K5^{hi}-sorted compared with K5^{low}-sorted populations in both tumours analysed by EM-seq, with higher variability among the K5^{hi} 200-cell replicates (Supplementary Fig. 6f,g). Together, our analyses across DNA methylation, chromatin accessibility and histone PTMs indicate that the basal-like subpopulation exhibts a noisy epigenetic landscape and increased transcriptional heterogeneity.

Single-cell distributions of H3K4 and K9 methylation

Our CyTOF analysis highlighted a global increase in several histone PTMs in the basal-like subpopulation, including H3K4 monomethylation and trimethylation (Fig. 2e). To explore the genomic distribution of these marks, we conducted single-nucleus CUT&Tag (snCUT&Tag) on three luminal tumours from our cohort[60]. Similar to the CyTOF data (Fig. 1b), stroma cells clustered together across patients, showing

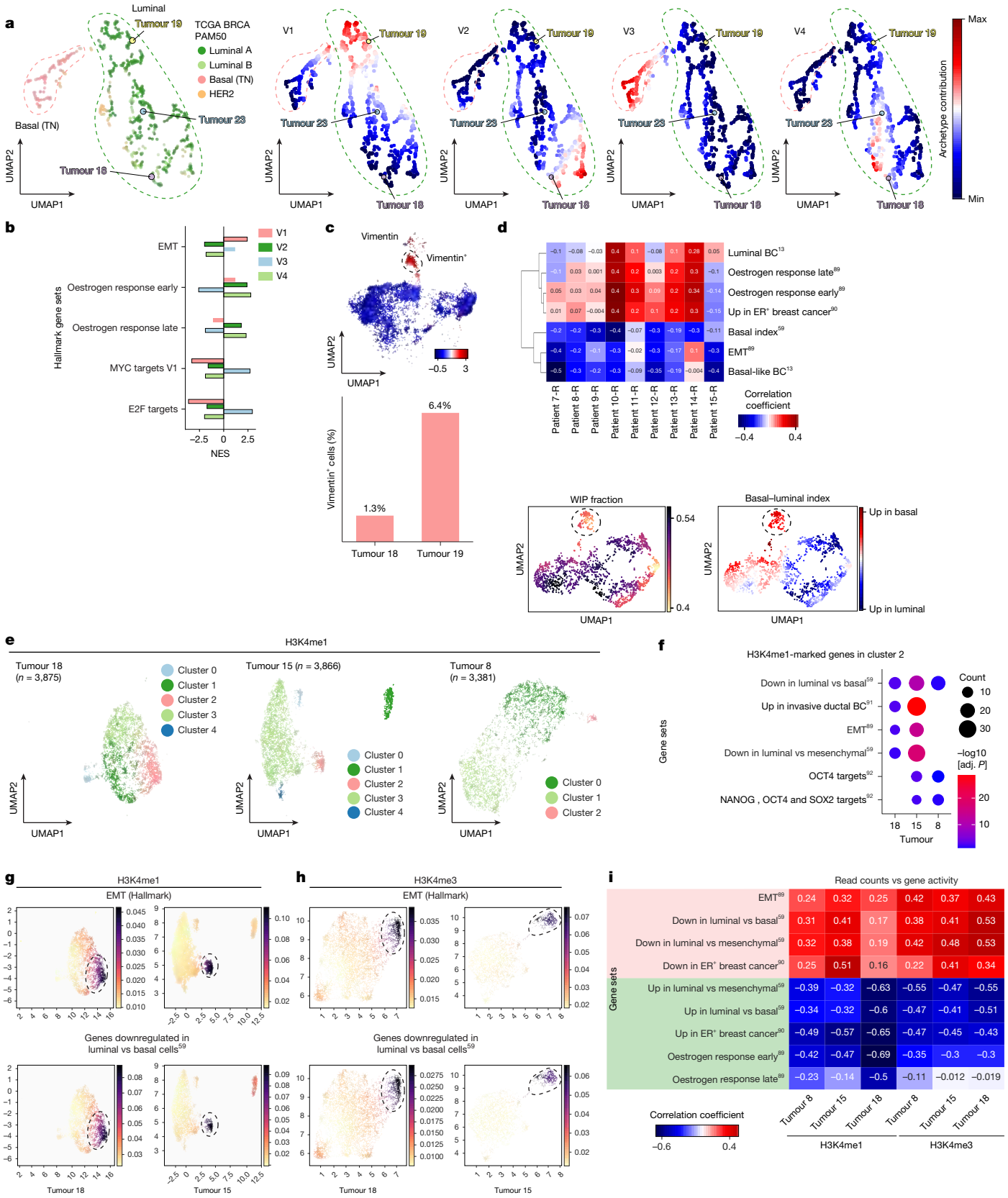


Fig. 3 | See next page for caption.

H3K4me1 and H3K4me3 deposition on fibroblast and endothelial markers. By contrast, tumours clustered separately, which indicated inter-patient variability (Extended Data Fig. 8a–d).

In each tumour, malignant cells clustered into distinct H3K4me1-defined or H3K4me3-defined states (Fig. 3e and Extended Data Fig. 8e,f). Differential analysis of H3K4me1 distribution between

Fig. 3 | DNA methylation and single-cell chromatin mapping support basal-like and luminal cell states. a, TCGA BRCA DNA-methylation dataset ($n = 782$)[51] deconvoluted into four archetypes, with UMAP of archetype composition per tumour coloured by PAM50 subtype (left) or archetype contribution (right). The three tumours (18, 19 and 23) analysed in this study are indicated and vary mainly along the V1 archetype. **b**, Genes ranked by correlation between DNA-methylation archetypes and matched TCGA BRCA expression data, followed by gene set enrichment analysis (GSEA) against the MSigDB hallmark collection. V1 differs from the luminal archetypes (V2 and V4) on the EMT axis. Representative normalized enrichment scores (NES) are shown. Black outline indicates Benjamini-Hochberg-adjusted $P < 0.05$. All terms provided in Supplementary Fig. 5b. **c**, Top, UMAP of CyTOF measurements (tumours 18 and 19, stroma excluded), coloured by vimentin expression, with vimentin-positive cells highlighted. Bottom, the vimentin-positive fraction per tumour. **d**, Top, scATAC–seq of luminal tumours ($n = 9$)[13] analysed for WIP fraction[58] and gene activity. Matrix shows Pearson's correlations between the WIP fraction and UCell scores per tumour. Bottom, UMAPs of cancer cells from patient 10-R, coloured by WIP fraction or the basal–luminal index (the difference between *z*-scored luminal and basal gene sets[59]; higher values indicate more basal-like expression). **e**, UMAP of H3K4me1 snCUT&Tag data (stroma removed). Cells are represented by a scTransform-normalized gene activity matrix reflecting H3K4me1-associated regulatory activity per gene, with the read count excluded. Louvain clustering and cell numbers per tumour are indicated. **f**, Differentially H3K4me1-marked genes per cluster in **e** (two-sided Wilcoxon signed-rank test, $\log_2$ fold change > 1, adjusted $P < 0.01$) were analysed for enrichment against MSigDB Hallmark and breast-related collections (Supplementary Table 4). Shown are terms enriched among genes with increased H3K4me1 in cluster 2 in at least two out of three tumours (one-sided hypergeometric test, FDR < 0.05). Circle sizes indicate overlapping genes; colour, $-\log_{10}$[FDR]. **g**,**h**, UMAPs of H3K4me1 (**g**) and H3K4me3 (**h**) snCUT&Tag data, analysed as in **e** and **f** and coloured by UCell scores for signatures comprising genes contributing to the respective enrichments (Gaussian-smoothed). For H3K4me3 enrichment, see Extended Data Fig. 8h. **i**, Pearson's correlations between snCUT&Tag read number and UCell scores for the indicated gene signatures per tumour and histone mark. Gene signatures shown in **d**,**f**,**i** are derived from previously published studies (refs. 13,59,85–88).

clusters highlighted a recurrent group of cells across tumours, annotated as 'cluster 2', with significant H3K4me1 enrichment at genes associated with basal-like identity and EMT (Fig. 3f,g and Extended Data Fig. 8g,i). In tumour 15, cluster 1 was also enriched for basal-like genes (adjusted $P < 0.05$), but not EMT. H3K4me3 similarly identified basal-like subpopulations in all tumours (Fig. 3h and Extended Data Fig. 8h,i). Accordingly, peaks associated with these clusters showed enrichment for transcription factor motifs linked to EMT and basal-like identity (Supplementary Fig. 7a). Next, we trained a classifier to identify H3K4me1-marked and H3K4me3-marked genes that contribute to variation between the clusters. For both histone methylations, these genes were enriched in terms related to breast cancer cellular identity, including oestrogen signalling, basal-like and mesenchymal identity (Supplementary Fig. 7b–e). Overall, our analysis provides support for chromatin-defined cell states in luminal tumour cells, including a basal-like state distinguished by H3K4me1 and H3K4me3 distributions.

The above snCUT&Tag results aligned with the heterogeneity observed by CyTOF, which led us to compare the single-cell datasets and integrating both genomic and protein information. CyTOF-defined basal-like cells showed global increases in both H3K4 methylation marks. We reasoned that sequencing reads per nucleus may reflect global levels of the target modification. Single-cell epigenomic analyses generally assume similar global modification levels, discarding read-count information, as in our analysis above[61] (Fig. 3e–h). Departing from this convention, we examined whether read counts capture additional aspects of epigenetic heterogeneity. Across all three tumours and both modifications, read counts negatively correlated with luminal identity signatures and positively with basal-like and mesenchymal signatures (Fig. 3i and Extended Data Fig. 9a,b). Consistently, cells with high read counts, potentially indicative of elevated global levels of H3K4me1 or H3K4me3, showed less enrichment of these active marks at luminal identity genes. Although requiring further validation, this direct comparison suggests that read counts may serve as proxy for global levels of histone modifications to capture additional aspects of epigenetic variation.

To further integrate these datasets, we label-transferred the CyTOF classification (Figs. 1 and 2) onto the H3K4me1 snCUT&Tag dataset. CyTOF-predicted basal-like populations were enriched in the basal-like clusters identified by H3K4me1 genomic profiling (Fig. 3e and Extended Data Fig. 9b). For tumour 15, cluster 1 was also predicted to comprise basal-like cells, consistent with its higher H3K4me1 deposition on basal-like-associated genes (Fig. 3g and Extended Data Fig. 9c). CyTOF-predicted basal-like cells also showed higher H3K4me1 deposition at EMT and basal-like gene signatures, which further confirmed their distinct epigenetic and functional features (Extended Data Fig. 9d). To validate these results, we sorted luminal PDX models and MCF7 cells by K5 expression, followed by H3K4me1 CUT&Tag. $K5^{low}$-specific peaks were enriched near oestrogen signalling genes, whereas $K5^{hi}$-specific peaks were associated with basal-like and mesenchymal programs, which confirmed the distinct H3K4me1 landscapes between subpopulations (Extended Data Fig. 9e). Together, the protein abundance and genomic distribution data implicate H3K4me1 and H3K4me3 as epigenetic determinants of the basal-like subpopulation in luminal tumours.

The repressive mark H3K9me2 was also identified as a strong contributor to the epigenetic basal-like signature (Fig. 2d,e). Thus, we conducted H3K9me2 snCUT&Tag on tumour 8, which was also profiled for H3K4me1. Given that H3K9me2 occupies large genomic regions, its distribution is not readily interpretable in relation to gene activity, especially in low-coverage single-cell datasets. We therefore leveraged the H3K4me1 data and trained a classifier, assuming opposing genomic distributions of enhancer-associated H3K4me1 and repressive H3K9me2 chromatin[62,63], to identify malignant versus stromal cells (Extended Data Fig. 9f). We next clustered tumour cells by H3K9me2 genomic distribution and applied the same H3K4me1-based inference framework. This approach predicted strong enrichment of basal-like cells in cluster 4, consistent with elevated H3K9me2 deposition on oestrogen-related genes (Extended Data Fig. 9g–i). This cluster also showed the highest H3K9me2 read counts per nucleus, which provided support for its correspondence to the $H3K9me2^{hi}$ basal-like cells identified by CyTOF (Extended Data Fig. 9j). Overall, these analyses highlight an epigenetically distinct basal-like population with globally increased and redistributed active and repressive histone modifications over genes associated with luminal-basal identity and EMT.

Single-molecule histone PTM profiling in blood

Our multimodal analysis linked the chromatin landscape to intratumoural cell-state diversity. We next aimed to translate our tissue-level findings to a liquid biopsy test based on our single-molecule EPINUC technology, which profiles the epigenetics of plasma-isolated nucleosomes[28,29]. We tailored EPINUC for breast cancer by developing single-molecule systems to quantify the breast-tissue-related proteins K5, L1CAM and MUC1, and added the histone PTMs H4K16ac and H4K20me3 (refs. 64–67) (Fig. 4a, Supplementary Table 5 and Supplementary Fig. 8a–c). Nucleosomes from plasma were captured and their combinatorial modifications imaged by TIRF microscopy, whereas protein biomarkers were quantified in parallel single-molecule antibody-capture assays (Fig. 4a). The expanded 29-parameter panel was applied to a cohort of plasma from 77 women with luminal breast cancer and 30 healthy women, and the results revealed significant differences (Supplementary Table 6). For example, cancer samples contained more H3K9ac and H4K16ac nucleosomes (Fig. 4b). As PTMs were normalized to nucleosome abundance, these differences reflect

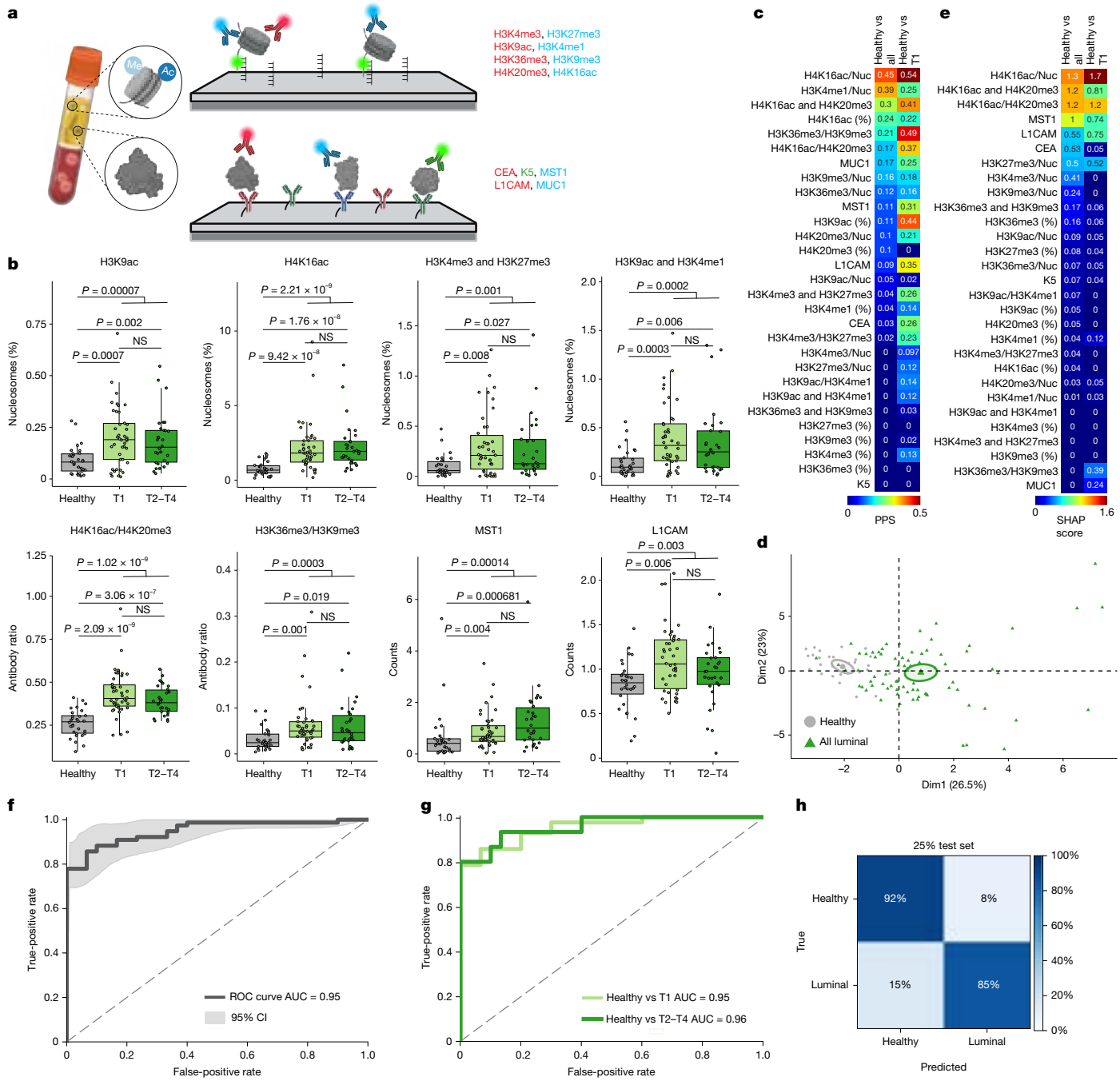


Fig. 4 | Expanded EPINUC analysis enables the detection of early-stage luminal breast cancer from blood samples. a, Schematic of the experiment. Circulating nucleosomes and proteins are captured on a microscope coverslip, which facilitates sensitive quantification of multiple histone PTMs and protein biomarkers from <1 ml plasma[28]. Fluorophore-conjugated antibodies enable simultaneous measurement of multiple marks, including their ratios and co-occurrence on individual nucleosomes. **b**, Selected EPINUC features in healthy women ($n = 30$) and in patients with stage T1 ($n = 42$) or stage T2–T4 luminal breast cancer ($n = 30$; Supplementary Tables 5 and 6). Boxes show median and 25–75th percentiles, and whiskers extend to the most extreme values within 1.5× the IQR. Circles represent individual samples. *P* values were calculated by two-sided two-sample Wilcoxon test and corrected for multiple comparisons (Bonferroni). NS, not significant. **c**, Individual PPS analysis distinguishing healthy individuals from all patients with luminal breast cancer or stage T1 breast cancer. Colour indicates the PPS values, reflecting the predictive power of each individual feature. See Supplementary Table 5 for EPINUC feature annotation. **d**, PCA of healthy ($n = 30$) and luminal breast cancer ($n = 77$) samples using all contributing features shown in **c** (PPS > 0). Samples are coloured by group; ellipses indicate 95% CI for the group barycentre. **e**, XGBoost classifier distinguishing healthy individuals from all patients with luminal breast cancer or stage T1 breast cancer, using all features measured and their combinations. SHAP scores indicate the contribution of each feature to classifier performance. **f**,**g**, Receiver operating characteristic (ROC) curves for random forest models discriminating healthy individuals ($n = 30$) from all patients with luminal breast cancer ($n = 77$; **f**), T1 luminal breast cancer ($n = 42$; **g**) or T2–T4 luminal breast cancer ($n = 30$; **g**). Models used multiplicative feature combinations followed by Boruta feature selection. AUCs were 0.95 (95% CI, 0.91–0.98), 0.95 (95% CI, 0.90–0.99) and 0.96 (95% CI, 0.91–1.00), respectively; shaded areas indicate 95% CI. **h**, Confusion matrix of model predictions for the 25% test set from the multi-institutional cohort (188 samples), comparing samples from healthy individuals ($n = 13$) and from patients with luminal breast cancer ($n = 34$), with specificity set to 0.9. Rows were normalized to show predictions as a percentage of samples in each true group. Schematic in **a** created in BioRender; Furth, N. https://biorender.com/qxlshkj (2026).

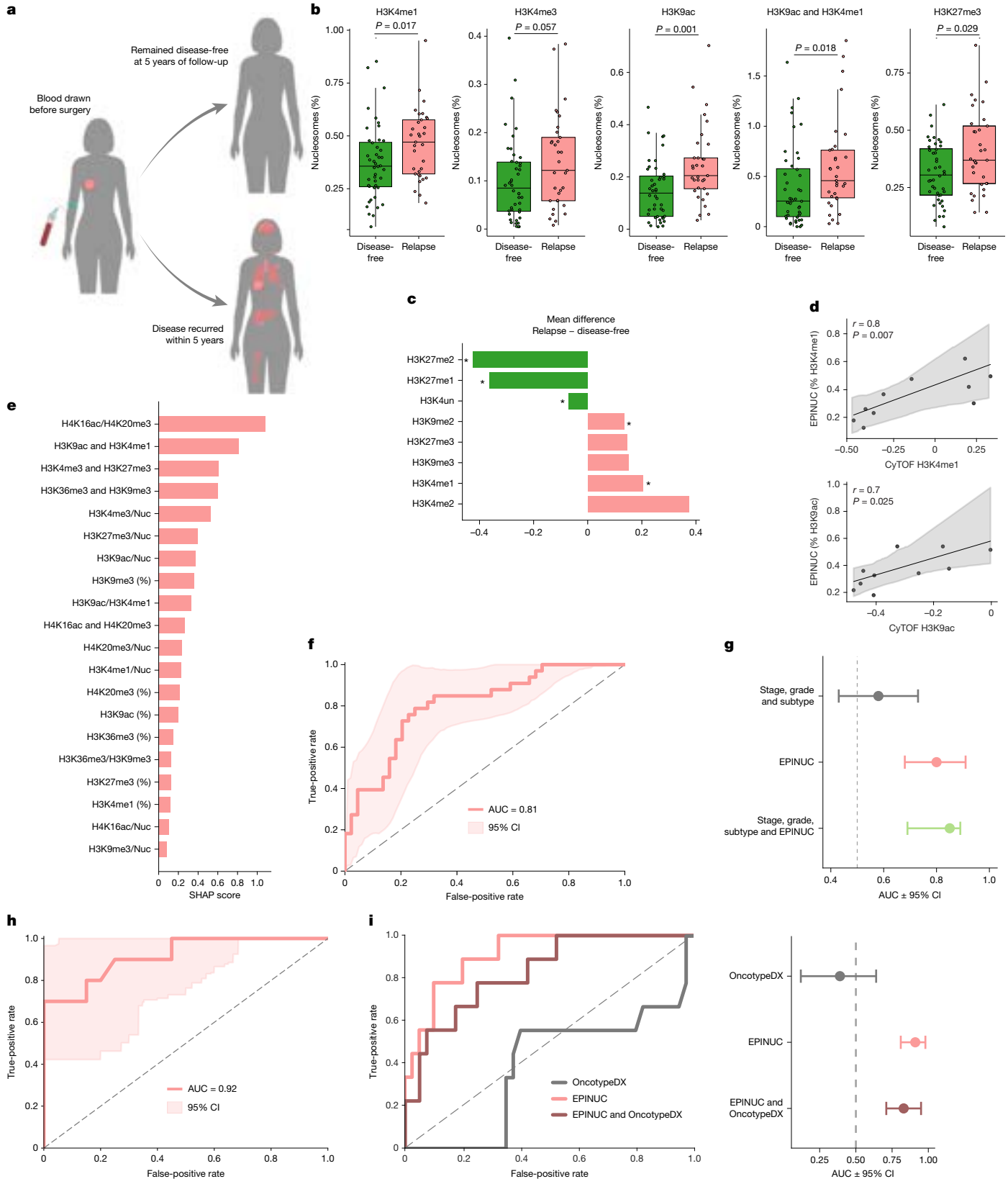


Fig. 5 | See next page for caption.

the epigenetic state rather than tumour burden. We also used EPINUC to measure PTM ratios and co-occurrence, which revealed increased bivalent (H3K4me3 and H3K27me3) nucleosomes, as previously observed in colorectal cancer and pancreatic cancer[28]. Indeed, bivalency is altered across cancers[68], and can prime drug-tolerant persister programs in breast cancer[19].

L1CAM and MST1 were also elevated in plasma from patients with breast cancer, a result in agreement with the association of high MST1

Fig. 5 | EPINUC reveals differences in PTMs associated with basal-like cells in the plasma of patients who developed recurrent disease. a, Schematic of the experiment. Blood samples were collected before surgery from patients with luminal breast cancer, with clinical follow-up of at least 5 years. Samples were stratified on the basis of disease outcome: disease-free ($n = 44$) or relapse ($n = 33$). **b**, Selected EPINUC features in the two outcome groups defined in **a**. Boxes show median and 25–75th percentiles, and whiskers extend to the most extreme values within 1.5× the IQR. Circles represent individual samples. Two-sided Wilcoxon test. **c**, Histone PTM-focused mass-spectrometry profiling of luminal tumours stratified by outcome (disease-free, $n = 17$; relapse, $n = 4$). Shown are differences in the mean levels of modifications relevant to the CyTOF or EPINUC panels between the two groups. Two-sided Welch's test; $^*P < 0.05$. Exact P values are provided in the source data. **d**, Scatter plot comparing mean histone PTM levels measured by CyTOF (tumour cells) with EPINUC measurements in matched plasma ($n = 10$). Spearman's correlation coefficient and two-sided P values are indicated. Lines show Theil–Sen fit, shading indicates 95% CI. **e**, XGBoost classifier distinguishing disease-free ($n = 44$) and relapse ($n = 33$) groups using all EPINUC features. SHAP scores indicate the contribution of each epigenetic feature to classifier performance, ranked by importance. **f**–**i**, Recurrence prediction in patients with luminal breast cancer. **f**, ROC curve for the MAYO-1 cohort (disease-free, $n = 44$; relapse, $n = 33$; AUC = 0.81, 95% CI 0.71–0.90). **g**, AUC estimates for patients with complete clinical annotation (disease-free, $n = 39$; relapse, $n = 28$), using clinical prognostic variables (age, tumour stage, lymph-node stage, grade and luminal subtype), EPINUC features or both. **h**, ROC curve for the independent MAYO-2 cohort (disease-free, $n = 20$; relapse, $n = 10$; AUC = 0.92, 95% CI 0.79–1.00). **i**, ROC curves and AUC estimates using OncotypeDX score, EPINUC features, or both, among patients with available OncotypeDX measurements, comparing disease-free ($n = 40$) and relapse ($n = 9$). For ROC curves, shading indicates 95% CI; for AUC estimates, points and lines indicate AUC and 95% CI, respectively. Schematic in **a** created in BioRender; Furth, N. https://biorender.com/xg9gtix (2026).

with poor prognosis in ER$^+$ tumours[69] (Fig. 4b and Supplementary Fig. 8d). Multiple parameters differed between healthy individuals and patients diagnosed with early-stage (T1) low-burden tumours, which demonstrated the early detection sensitivity of this method (Fig. 4b). Predictive power scores (PPS) revealed the contribution of the newly measured H4 PTMs, selected on the basis of the CyTOF analysis, which illustrates the value of matched tissue–plasma profiling (Fig. 4c). Principal component analysis (PCA) of predictive features separated healthy individuals from patients with breast cancer, which highlights the ability of EPINUC to capture differences between these groups (Fig. 4d).

We next applied machine-learning-based classifiers to diagnose luminal breast cancer from EPINUC data. SHAP analysis of XGBoost models trained on all patients, or only stage T1 cases, identified multiple histone PTMs and protein biomarkers as informative (Fig. 4e). Notably, both models emphasized combinatorial features, particularly H4K16ac and H4K20me3 co-occurrence, thereby highlighting the ability of EPINUC to quantify such combinations. In particular, H4K16ac activates transcription of transposable elements[70], which reside in H4K20me3 heterochromatin[71], and because transposable elements are often upregulated in cancer, elevated H4K16ac and H4K20me3 might be a cancer-associated phenomenon. Finally, we applied several classifiers to assess the diagnostic contribution of the expanded EPINUC panel in discriminating healthy individuals from patients with cancer. Use the original 20 EPINUC parameters[28] produced an area under the curve (AUC) of 0.85 (Supplementary Fig. 9a), whereas the expanded 29-parameter panel achieved 0.95, including among patients with early-stage T1 cancer (Fig. 4f,g). Unlike mammography, the sensitivity of which declines with dense breast tissue[72,73], the features and performance of EPINUC were unaffected by different BI-RADS breast density scores, including patients with heterogeneous and extremely dense breast scores (Supplementary Fig. 9b,c). To assess the robustness of the EPINUC-based classifier, we applied the model trained on the original cohort (MAYO-1; Fig. 4f) to an independent, separately processed set of plasma samples collected 6–12 months later (MAYO-2). Following batch correction (Methods), the pre-trained model accurately distinguished healthy from luminal breast cancer, achieving an AUC of 0.98 and high classification accuracy (Extended Data Fig. 10a,b).

To generalize our findings, we analysed pretreatment plasma from patients with luminal breast cancer at two independent medical centres (Supplementary Table 6). EPINUC distinguished patients from healthy controls with high accuracy in both cohorts and showed comparable sensitivity and specificity in TNBC (Extended Data Fig. 10c–e). Pooling all luminal samples across centres ($n = 188$), a model trained on 75% of the data achieved an AUC of 0.96 and robustly classified the remaining 25% despite mixed sources and batches, a scenario that reflects real-world clinical settings (Fig. 4h, Extended Data Fig. 10f and Supplementary Fig. 10). Performance was similar for luminal A and luminal B tumours (Extended Data Fig. 10g). Although retrospective, these results support the use of high-resolution, multiparametric plasma epigenetic analyses for cancer diagnosis.

MAYO-1 comprised preoperative samples with at least 5 years of clinical follow-up, which enabled stratification by 5-year recurrence (Fig. 5a). Already at diagnosis, several EPINUC features differed between patients who were disease-free and patients who had relapsed disease. H3K4 monomethylation and trimethylation, associated with the aggressive basal-like subpopulation, were elevated in plasma of patients developing recurrent disease (Fig. 5b). H3K27me3 and H3K9ac were also elevated, perhaps reflecting other aggressive tumour or microenvironmental states, such as cycling cells that were particularly high for histone acetylations (Extended Data Fig. 1b). Nucleosomes modified with both H3K9ac and H3K4me1 were higher in the recurrence group, a result that further demonstrates the power of EPINUC in capturing these combinations (Fig. 5b). ER or PR status did not associate with relapse (Fisher's exact test, $P > 0.05$), and the epigenetic signature persisted in tumours with the following characteristics: comparable ER and PR levels; only luminal A tumours; low-burden T1–T2 or grade 1–2 disease; and tumours without lymphovascular invasion (LVI). No EPINUC feature was significantly associated with histological grade, which indicates that there is prognostic information beyond known clinical parameters (Extended Data Fig. 11a–f).

To further explore the association between tumour and plasma epigenetic patterns, we analysed a small cohort of luminal tumours with histone PTM mass-spectrometry data[39], which were similarly stratified by disease recurrence[39]. H3K4me1 and H3K9me2, which marked the basal-like cells in our CyTOF data, were significantly higher in patients who developed recurrence, a result that further links this subpopulation to aggressive tumour traits (Figs. 2e and 5c). Finally, for patients analysed in Figs. 1–3, we obtained preoperative blood samples, which enabled direct correlation of tumour (CyTOF) and blood (EPINUC) epigenetic patterns. Most histone PTMs correlated positively, with the recurrence-associated H3K4me1 and H3K9ac marks showing the strongest positive correlation, a result that provided further support for their prognostic value (Fig. 5d and Extended Data Fig. 11h). Although requiring larger validation, these data suggest that circulating nucleosomes report both the tumour presence and its epigenetic state, a finding with potentially important implications for clinical management.

To examine the prognostic power of EPINUC to predict disease recurrence, we applied machine-learning classifiers on the multiparametric data. The best model achieved an AUC of 0.81, and outperformed predictions based on classical prognostic parameters such as stage, grade and subtype (Fig. 5e–g and Extended Data Fig. 11i). To assess reproducibility and to benchmark established assays, MAYO-2 was focused on samples with available OncotypeDX scores[74]. Despite its smaller size, EPINUC predicted recurrence with an AUC of 0.92 (Fig. 5h). Across patients with OncotypeDX data from both cohorts, scores did not differ between disease-free and relapsed groups, whereas EPINUC

retained discriminatory power (Fig. 5i, Extended Data Fig. 11j and Supplementary Fig. 11). Combining assays did not improve performance, which indicated limited added value from OncotypeDX in this setting. Overall, the data suggest that epigenetic signatures of tumour cell subpopulations, such as basal-like cells, may be captured in blood and provide insight into disease biology beyond the static tissue-based prognostic window. Pending larger prospective validation, these findings provide support for the use of EPINUC for early diagnosis and recurrence-risk stratification.

Discussion

Single-cell studies have revealed clinically relevant phenotypic diversity across cancers, yet the epigenetic mechanisms fuelling this heterogeneity remain poorly understood[75,76]. Our analysis of luminal breast tumours uncovered tumour cell diversity consistent with previous transcriptomic and proteomic approaches[14,21,33,34,54]. Quantitative mass-spectrometry also showed differences in global histone modification levels between breast cancer subtypes, which provides support for the idea that chromatin regulates cell identity[39,77,78]. Altered H3K4 demethylation can drive loss of luminal identity[40], which suggests that changes in ER signalling, even without *ESR1* mutations, may promote luminal-to-basal plasticity. Here we performed single-cell profiling of histone PTMs in luminal breast tumours. By measuring multiple global chromatin modifications at single-cell resolution, we identified a recurrent epigenetic signature of $K5^{hi}$ basal-like cells across tumours, supported across modalities by single-cell genomic profiling of H3K4 methylation.

Reconstitution of the full heterogeneity spectrum in single-cell-derived MCF7 clones and independently grown PDX tumours provides support for epigenetically driven, interconverting states rather than genetically distinct clones. Subpopulation frequencies varied among patients and may be influenced by mutations that bias state acquisition, as reported in glioblastoma[46]. This plasticity may enable adaptation to therapeutic pressures, which then promotes resistance and progression. Accordingly, proteomic intratumoural heterogeneity increases with tumour grade independently of genomic heterogeneity, thereby implicating epigenetic mechanisms[7]. Whether epigenetic therapies can limit plasticity and state transitions remains to be tested.

Breast cancer prognosis depends strongly on early detection, yet accurate screening while minimizing false positives remains challenging[79–82]. Detection of molecular changes before clinical progression could guide treatment adaptation and maximize first-line benefit[83]. Tumour-level histone PTM signatures distinguished luminal breast cancer cell states, whereas EPINUC analysis of pretreatment plasma samples provided proof-of-concept for outcome prediction. Notably, H3K4 methylation marks were elevated in both basal-like tumour cells and in plasma from patients with recurrent disease, which suggests that there is a link between this subpopulation and disease outcome.

Unlike CyTOF, which resolves tumour cell populations, plasma measurements integrate signals from damage of healthy tissue, the tumour microenvironment and local and systemic immune responses[29,84]. Nevertheless, patient-matched histone PTMs analyses in tumours and blood revealed a positive correlation for most marks, with the highest correlation observed for the recurrence-associated marks H3K4me1 and H3K9ac, which provides support for their prognostic value. Overall, our tissue and plasma chromatin dataset reveals epigenetically driven intratumoural heterogeneity in breast cancer and supports the use of EPINUC as a potential approach for breast cancer diagnosis and prognosis.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-026-11230-y.

1. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2022. *CA Cancer J. Clin.* **72**, 7–33 https://doi.org/10.3322/caac.21708 (2022).
2. Perou, C. M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 https://doi.org/10.1038/35021093 (2000).
3. Sørlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98**, 10869–10874 https://doi.org/10.1073/pnas.191367098 (2001).
4. Laisné, M., Lupien, M. & Vallot, C. Epigenomic heterogeneity as a source of tumour evolution. *Nat. Rev. Cancer* **25**, 7–26 https://doi.org/10.1038/s41568-024-00757-9 (2025).
5. Mazor, T., Pankov, A., Song, J. S. & Costello, J. F. Intratumoral heterogeneity of the epigenome. *Cancer Cell* **29**, 440–451 https://doi.org/10.1016/j.ccell.2016.03.009 (2016).
6. Baslan, T. et al. Novel insights into breast cancer copy number genetic heterogeneity revealed by single-cell genome sequencing. *eLife* **9**, e51480 https://doi.org/10.7554/eLife.51480 (2020).
7. Mardamshina, M. et al. Integrated spatial proteomic analysis of breast cancer heterogeneity unravels cancer cell phenotypic plasticity. *Nat. Commun.* **16**, 10482 https://doi.org/10.1038/s41467-025-65477-6 (2025).
8. Rueda, O. M. et al. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature* **567**, 399–404 https://doi.org/10.1038/s41586-019-1007-8 (2019).
9. Yeo, S. K. & Guan, J. L. Breast cancer: multiple subtypes within a tumor? *Trends Cancer* **3**, 753–760 https://doi.org/10.1016/j.trecan.2017.09.001 (2017).
10. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).
11. Rivera, C. M. & Ren, B. Mapping human epigenomes. *Cell* **155**, 39–55 https://doi.org/10.1016/j.cell.2013.09.011 (2013).
12. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 https://doi.org/10.1038/47412 (2000).
13. Regner, M. J. et al. Defining the regulatory logic of breast cancer using single-cell epigenetic and transcriptome profiling. *Cell Genomics* **5**, 100765 https://doi.org/10.1016/j.xgen.2025.100765 (2025).
14. Wu, S. Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 https://doi.org/10.1038/s41588-021-00911-1 (2021).
15. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 https://doi.org/10.1038/nature14590 (2015).
16. Iglesia, M. D. et al. Differential chromatin accessibility and transcriptional dynamics define breast cancer subtypes and their lineages. *Nat. Cancer* **5**, 1713–1736 https://doi.org/10.1038/s43018-024-00773-6 (2024).
17. Alcolea, M. P. et al. Cancer hallmarks: piecing the puzzle together. *Cancer Discov.* **14**, 674–682 https://doi.org/10.1158/2159-8290.Cd-24-0097 (2024).
18. Grosselin, K. et al. High-throughput single-cell ChIP–seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* **51**, 1060–1066 https://doi.org/10.1038/s41588-019-0424-9 (2019).
19. Marsolier, J. et al. H3K27me3 conditions chemotolerance in triple-negative breast cancer. *Nat. Genet.* **54**, 459–468 https://doi.org/10.1038/s41588-022-01047-6 (2022).
20. Harpaz, N. et al. Single-cell epigenetic analysis reveals principles of chromatin states in H3.3-K27M gliomas. *Mol. Cell* **82**, 2696–2713 https://doi.org/10.1016/j.molcel.2022.05.023 (2022).
21. Aylon, Y. et al. Breast cancer plasticity is restricted by a LATS1–NCOR1 repressive axis. *Nat. Commun.* **13**, 7199 https://doi.org/10.1038/s41467-022-34863-9 (2022).
22. Furth, N. et al. Oncogenic $IDH1^{mut}$ drives robust loss of histone acetylation and increases chromatin heterogeneity. *Proc. Natl Acad. Sci. USA* **122**, e2403862122 https://doi.org/10.1073/pnas.2403862122 (2025).
23. Griess, O. et al. Mutant EZH2 alters the epigenetic network and increases epigenetic heterogeneity in B cell lymphoma. *PLoS Biol.* **23**, e3003191 https://doi.org/10.1371/journal.pbio.3003191 (2025).
24. Lui, Y. Y. et al. Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. *Clin. Chem.* **48**, 421–427 (2002).
25. van der Pol, Y. & Mouliere, F. Toward the early detection of cancer by decoding the epigenetic and environmental fingerprints of cell-free DNA. *Cancer Cell* **36**, 350–368 https://doi.org/10.1016/j.ccell.2019.09.003 (2019).
26. Gezer, U. & Holdenrieder, S. Post-translational histone modifications in circulating nucleosomes as new biomarkers in colorectal cancer. *In Vivo* **28**, 287–292 (2014).
27. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* **164**, 57–68 https://doi.org/10.1016/j.cell.2015.11.050 (2016).
28. Fedyuk, V. et al. Multiplexed, single-molecule, epigenetic analysis of plasma-isolated nucleosomes for cancer diagnostics. *Nat. Biotechnol.* **41**, 212–221 https://doi.org/10.1038/s41587-022-01447-3 (2023).
29. Erez, N. et al. Single-molecule systems for the detection and monitoring of plasma-circulating nucleosomes and oncoproteins in diffuse midline glioma. *Cell Rep. Med.* **6**, 101918 (2025).
30. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 https://doi.org/10.1038/s42256-019-0138-9 (2020).
31. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proc. 31st International Conference on Neural Information Processing Systems* 4768–4777 (Curran Associates, 2017).
32. Nielsen, T. O. et al. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clin. Cancer Res.* **16**, 5222–5232 https://doi.org/10.1158/1078-0432.Ccr-10-1282 (2010).

33. Ali, H. R. et al. Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat. Cancer* **1**, 163–175 https://doi.org/10.1038/s43018-020-0026-6 (2020).
34. Georgopoulou, D. et al. Landscapes of cellular phenotypic diversity in breast cancer xenografts and their impact on drug response. *Nat. Commun.* **12**, 1998 https://doi.org/10.1038/s41467-021-22303-z (2021).
35. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016).
36. Parreno, V., Martinez, A.-M. & Cavalli, G. Mechanisms of Polycomb group protein function in cancer. *Cell Res.* **32**, 231–253 https://doi.org/10.1038/s41422-021-00606-6 (2022).
37. Chang, C.-J. et al. EZH2 promotes expansion of breast tumor initiating cells through activation of RAF1–β-catenin signaling. *Cancer Cell* **19**, 86–100 https://doi.org/10.1016/j.ccr.2010.10.035 (2011).
38. Chan, H. L. & Morey, L. Emerging roles for Polycomb-group proteins in stem cells and cancer. *Trends Biochem. Sci.* **44**, 688–700 https://doi.org/10.1016/j.tibs.2019.04.005 (2019).
39. Noberini, R. et al. A histone-centric multi-omics study shows that increased H3K4 methylation sustains triple-negative breast cancer phenotypes. *Nat. Commun.* **16**, 8716 https://doi.org/10.1038/s41467-025-63745-z (2025).
40. Yamamoto, S. et al. *JARID1B* is a luminal lineage-driving oncogene in breast cancer. *Cancer Cell* **25**, 762–777 https://doi.org/10.1016/j.ccr.2014.04.024 (2014).
41. Haque, R. et al. Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades. *Cancer Epidemiol. Biomarkers Prev.* **21**, 1848–1855 https://doi.org/10.1158/1055-9965.Epi-12-0474 (2012).
42. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 https://doi.org/10.1200/jco.2008.18.1370 (2009).
43. Prat, A. & Perou, C. M. Deconstructing the molecular portraits of breast cancer. *Mol. Oncol.* **5**, 5–23 https://doi.org/10.1016/j.molonc.2010.11.003 (2011).
44. Barroso-Sousa, R. et al. Prevalence and mutational determinants of high tumor mutation burden in breast cancer. *Ann. Oncol.* **31**, 387–394 https://doi.org/10.1016/j.annonc.2019.11.010 (2020).
45. Haricharan, S., Bainbridge, M. N., Scheet, P. & Brown, P. H. Somatic mutation load of estrogen receptor-positive breast tumors predicts overall survival: an analysis of genome sequence data. *Breast Cancer Res. Treat.* **146**, 211–220 https://doi.org/10.1007/s10549-014-2991-x (2014).
46. Neftel, C. et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* **178**, 835–849 https://doi.org/10.1016/j.cell.2019.06.024 (2019).
47. Kwon, T. et al. Mechanism of histone lysine methyl transfer revealed by the structure of SET7/9–AdoMet. *EMBO J.* **22**, 292–303 https://doi.org/10.1093/emboj/cdg025 (2003).
48. Xiao, B. et al. Structure and catalytic mechanism of the human histone methyltransferase SET7/9. *Nature* **421**, 652–656 https://doi.org/10.1038/nature01378 (2003).
49. Black, J. C., Van Rechem, C. & Whetstine, J. R. Histone lysine methylation dynamics: establishment, regulation, and biological impact. *Mol. Cell* **48**, 491–507 https://doi.org/10.1016/j.molcel.2012.11.006 (2012).
50. Gray, G. K. et al. A human breast atlas integrating single-cell proteomics and transcriptomics. *Dev. Cell* **57**, 1400–1420 https://doi.org/10.1016/j.devcel.2022.05.003 (2022).
51. Koboldt, D. C. et al. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 https://doi.org/10.1038/nature11412 (2012).
52. Sarrió, D. et al. Epithelial–mesenchymal transition in breast cancer relates to the basal-like phenotype. *Cancer Res.* **68**, 989–997 https://doi.org/10.1158/0008-5472.Can-07-2017 (2008).
53. Aleskandarany, M. A. et al. Epithelial mesenchymal transition in early invasive breast cancer: an immunohistochemical and reverse phase protein array study. *Breast Cancer Res. Treat.* **145**, 339–348 https://doi.org/10.1007/s10549-014-2927-5 (2014).
54. Chung, W. et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* **8**, 15081 https://doi.org/10.1038/ncomms15081 (2017).
55. Scott, C. A. et al. Identification of cell type-specific methylation signals in bulk whole genome bisulfite sequencing data. *Genome Biol.* **21**, 156 https://doi.org/10.1186/s13059-020-02065-5 (2020).
56. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 https://doi.org/10.1038/nrg3230 (2012).
57. Riobo-Del Galdo, N. A., Lara Montero, Á & Wertheimer, E. V. Role of Hedgehog signaling in breast cancer: pathogenesis and therapeutics. *Cells* **8**, 375 https://doi.org/10.3390/cells8040375 (2019).
58. Gamble, N. et al. Thymic epithelial cells amplify epigenetic noise to promote immune tolerance. *Nature* **646**, 724–733 https://doi.org/10.1038/s41586-025-09424-x (2025).
59. Charafe-Jauffret, E. et al. Gene expression profiling of breast cell lines identifies potential new basal markers. *Oncogene* **25**, 2273–2284 https://doi.org/10.1038/sj.onc.1209254 (2006).
60. Wu, S. J. et al. Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. *Nat. Biotechnol.* **39**, 819–824 https://doi.org/10.1038/s41587-021-00865-z (2021).
61. Raimundo, F., Prompsy, P., Vert, J.-P. & Vallot, C. A benchmark of computational pipelines for single-cell histone modification data. *Genome Biol.* **24**, 143 https://doi.org/10.1186/s13059-023-02981-2 (2023).
62. Zylicz, J. J. et al. Chromatin dynamics and the role of G9a in gene regulation and enhancer silencing during early mouse development. *eLife* **4**, e09571 https://doi.org/10.7554/eLife.09571 (2015).
63. Wen, B., Wu, H., Shinkai, Y., Irizarry, R. A. & Feinberg, A. P. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nat. Genet.* **41**, 246–250 https://doi.org/10.1038/ng.297 (2009).
64. Moisini, I., Zhang, H., D'Aguiar, M., Hicks, D. G. & Turner, B. M. L1CAM expression in recurrent estrogen positive/HER2 negative breast cancer: a novel biomarker worth considering. *Appl. Immunohistochem. Mol. Morphol.* **29**, 287–292 https://doi.org/10.1097/pai.0000000000000909 (2021).
65. Doberstein, K. et al. L1CAM is expressed in triple-negative breast cancers and is inversely correlated with androgen receptor. *BMC Cancer* **14**, 958 https://doi.org/10.1186/1471-2407-14-958 (2014).
66. Li, J. et al. Tumor markers CA15-3, CA125, CEA and breast cancer survival by molecular subtype: a cohort study. *Breast Cancer* **27**, 621–630 https://doi.org/10.1007/s12282-020-01058-3 (2020).
67. Wang, B. et al. Combined low levels of H4K16ac and H4K20me3 predicts poor prognosis in breast cancer. *Int. J. Clin. Oncol.* **28**, 1147–1157 https://doi.org/10.1007/s10147-023-02378-y (2023).
68. Brown, R., Curry, E., Magnani, L., Wilhelm-Benartzi, C. S. & Borley, J. Poised epigenetic states and acquired drug resistance in cancer. *Nat. Rev. Cancer* **14**, 747–753 https://doi.org/10.1038/nrc3819 (2014).
69. Li, X. et al. YAP inhibits ERα and ER^+ breast cancer growth by disrupting a TEAD–ERα signaling axis. *Nat. Commun.* **13**, 3075 https://doi.org/10.1038/s41467-022-30831-5 (2022).
70. Pal, D. et al. H4K16ac activates the transcription of transposable elements and contributes to their *cis*-regulatory function. *Nat. Struct. Mol. Biol.* **30**, 935–947 https://doi.org/10.1038/s41594-023-01016-5 (2023).
71. He, J. et al. Transposable elements are regulated by context-specific patterns of chromatin marks in mouse embryonic stem cells. *Nat. Commun.* **10**, 34 https://doi.org/10.1038/s41467-018-08006-y (2019).
72. Boyd, N. F. et al. Mammographic density and the risk and detection of breast cancer. *N. Engl. J. Med.* **356**, 227–236 https://doi.org/10.1056/NEJMoa062790 (2007).
73. Kerlikowske, K. et al. Identifying women with dense breasts at high risk for interval cancer: a cohort study. *Ann. Intern. Med.* **162**, 673–681 https://doi.org/10.7326/m14-1465 (2015).
74. Paik, S. et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 https://doi.org/10.1056/NEJMoa041588 (2004).
75. Boxer, E. et al. Emerging clinical applications of single-cell RNA sequencing in oncology. *Nat. Rev. Clin. Oncol.* **22**, 315–326 https://doi.org/10.1038/s41571-025-01003-3 (2025).
76. McQuerry, J. A., Chang, J. T., Bowtell, D. D. L., Cohen, A. & Bild, A. H. Mechanisms and clinical implications of tumor heterogeneity and convergence on recurrent phenotypes. *J. Mol. Med.* **95**, 1167–1178 https://doi.org/10.1007/s00109-017-1587-4 (2017).
77. Restellini, C. et al. Alternative digestion approaches improve histone modification mapping by mass spectrometry in clinical samples. *Proteomics Clin. Appl.* **13**, e1700166 https://doi.org/10.1002/prca.201700166 (2019).
78. Noberini, R., Uggetti, A., Pruneri, G., Minucci, S. & Bonaldi, T. Pathology tissue-quantitative mass spectrometry analysis to profile histone post-translational modification patterns in patient samples. *Mol. Cell. Proteomics* **15**, 866–877 https://doi.org/10.1074/mcp.M115.054510 (2016).
79. Siu, A. L. et al. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. *Ann. Intern. Med.* **164**, 279–296 https://doi.org/10.7326/m15-2886 (2016).
80. Elezaby, M. et al. *BRCA* mutation carriers: breast and ovarian cancer screening guidelines and imaging considerations. *Radiology* **291**, 554–569 https://doi.org/10.1148/radiol.2019181814 (2019).
81. Anaby, D. et al. 'Earlier than early' detection of breast cancer in Israeli *BRCA* mutation carriers applying AI-based analysis to consecutive MRI scans. *Cancers* **15**, 3120 (2023).
82. Drukteinis, J. S., Mooney, B. P., Flowers, C. I. & Gatenby, R. A. Beyond mammography: new frontiers in breast cancer screening. *Am. J. Med.* **126**, 472–479 https://doi.org/10.1016/j.amjmed.2012.11.025 (2013).
83. Bidard, F.-C. et al. First-line camizestrant for emerging *ESR1*-mutated advanced breast cancer. *N. Engl. J. Med.* **393**, 569–580 https://doi.org/10.1056/NEJMoa2502929 (2025).
84. Van den Ackerveken, P. et al. A novel proteomics approach to epigenetic profiling of circulating nucleosomes. *Sci. Rep.* **11**, 7256 https://doi.org/10.1038/s41598-021-86630-3 (2021).
85. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 https://doi.org/10.1016/j.cels.2015.12.004 (2015).
86. van 't Veer, L. et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
87. Schuetz, C. S. et al. Progression-specific genes identified by expression profiling of matched ductal carcinomas in situ and invasive breast tumors, combining laser capture microdissection and oligonucleotide microarray analysis. *Cancer Res.* **66**, 5278–5286 (2006).
88. Ben-Porath, I. et al. An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. *Nat. Genet.* **40**, 499–507 (2008).

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026

Methods

Human participants

Collection and analyses of breast cancer tissue samples and matching blood samples were approved by the Israeli Ministry of Health Ethics Committees (Helsinki applications KMC-0028-22), followed by approval of the Weizmann Institutional Review Board (IRB, 2314-3).

The breast cancer samples for histone PTM mass-spectrometry-based analysis, which comprise 21 tumours, were obtained from patients undergoing surgery at the European Institute of Oncology and are part of a study approved by the institute's Ethical Committee (study UID 2550).

Plasma samples and data for luminal breast cancer patients were obtained from Mayo Clinic (IRB approval number 23-011311). Collection of plasma samples from breast cancer patients treated at the Sheba Medical Center was approved by the Institutional Review Board (IRB approval 7509-09). Plasma samples from healthy individuals were obtained from the Weizmann Institute of Science (IRB approval number 1850-1).

All studies were performed in accordance with the relevant guidelines and regulations, and informed consent was obtained from all participants. No direct identifying participant information is reported in this study. Participants' information is detailed in Supplementary Tables 2 and 6. No statistical methods were used to predetermine sample sizes. No randomization or blinding was performed in the collection, processing, or analysis of human samples.

PDX models

All animal studies were approved by the Weizmann Institute's Institutional Animal Care and Use Committee (IACUC), and they were performed in accordance with Israeli law. Mice were housed and handled in a pathogen-free, temperature-controlled (22 °C ± 1 °C) facility on a 12-h light–dark cycle. Animals were fed a regular chow diet (2018 Teklad Global 18% protein rodent diet) and given ad libitum access to food and water.

The TM00112 (PDX-1) and TM00386 (PDX-2) models were purchased from the Jackson Laboratory and implanted in 10-week-old female NSG mice (Jackson Laboratory). In brief, following euthanasia, tumours were removed from donor mice and cut into small fragments. A small pouch was made in the lower back of each mouse, and one fragment was inserted into the pouch. The wound was closed using a surgical clip. Clips were removed 4–5 days after surgery. Tumours were measured twice a week with a caliper, and body weight was measured once a week. The IACUC-approved maximum tumour size was 1,000 mm^3. Mice were euthanized when tumours reached approximately 500 mm^3, and tumours were collected for analyses. No statistical method was used to predetermine sample sizes. Animals were not randomized, and investigators were not blinded during the experiments.

Cell culture

MCF7 and MDA-MB-468 cell lines were obtained from the Tissue Culture Cell repository at the Weizmann Institute of Science. HCC70, HCC1937 and SUM149 cell lines were provided by I. Ben-Porath. HCC70 and HCC1937 cells were cultured in RPMI-1640 supplemented with 10% FBS, 1% L-glutamine and 1% penicillin–streptomycin. SUM149 cells were maintained in Ham's F-12 medium supplemented with 5% FBS, 1 μg ml^{-1} hydrocortisone (Sigma-Aldrich, H0888), 0.5 μg ml^{-1} insulin (Sigma-Aldrich, I5500), L-glutamine and penicillin–streptomycin. MDA-MB-468 and MCF7 cells were cultured in DMEM (Thermo Fisher Scientific, 41965-039) supplemented with 10% FBS (Sigma-Aldrich, F7524), 1% L-glutamine (Sartorius, 03-020-1B) and 1% penicillin–streptomycin (Corning, 30-002-CI). All cells were maintained at 37 °C in a humidified atmosphere containing 5% CO_2.

Single-cell-derived MCF7 clones and matching parental MCF7 cells were either provided by U. Ben-David[89] or generated in-house by single-cell sorting into 96-well plates using a Sony SH800S cell sorter (Sony Biotechnology) and expanded for approximately 4 months before experiments.

Cell lines were not independently authenticated but were routinely tested for mycoplasma contamination and were negative.

Tumour sample selection for different applications

Owing to the limited number of dissociated cells available from each tumour sample provided for research, we were unable to profile identical samples using all the different methods described in the paper. The single-cell CyTOF-based epigenetic analysis of human tumours was the priority for each tumour (besides tumour 23, which was used for WGBS and not profiled by CyTOF). For tumour 18, both WGBS and snCUT&Tag were performed owing to the high availability of tumour cells. For the remaining samples, we did not have sufficient cells to perform both methodologies. Notably, although the additional analyses were not conducted on identical samples, each was performed on three independent tumour samples, which ensured the reproducibility of the conclusions from each approach. The focus of the analysis was on identifying common patterns in tumours, rather than comparing between tumours and patients.

Clinicopathological and molecular data of tumour samples

Patient and tumour information was extracted from pathological reports. To quantify ER positivity, formalin-fixed paraffin-embedded (FFPE) tumour blocks prepared alongside sample collection for CyTOF analysis were retrieved, sectioned and stained for ER by IHC. Slides were scanned, and digital images containing DAB signals were analysed using QuPath software[90]. Tissue regions were first segmented using a pixel classifier to identify areas containing tissue (average channels < 245, sigma = 2, at 8.4 μm resolution). In these regions, epithelial and stromal compartments were segmented using a supervised pixel classifier, with a minimum object size of 1,000 $μm^2$ and a minimum hole size of 5,000 $μm^2$. The classifier was trained on representative regions from multiple slides using a multilayer perceptron (ANN_MLP) model implemented in QuPath, based on haematoxylin and DAB channels at 2.1 μm resolution and multiscale neighbourhood features (scales 1, 2, 4, 8 and 12). Cell nuclei in epithelial regions were segmented using Cellpose-SAM[91]. For each nucleus, the median DAB intensity was quantified. The distribution of nuclear DAB intensities was computed for each tumour sample and a per-sample thresholding was used to account for inter-slide staining variability. For each tumour, a kernel density estimate of the median nuclear DAB intensity was computed, constrained to the observed intensity range. Local maxima and minima of the density function were identified, and the threshold separating ER-negative and ER-positive nuclei was defined as the lowest local minimum between the two most prominent density peaks. Tumour-specific thresholds derived using this approach were applied to classify nuclei as ER-positive or ER-negative, and the percentage of ER-positive cells was calculated for each sample.

Mutation calling in tumour samples. DNA was extracted from FFPE blocks using a RecoverAll Total Nucleic Acid Isolation kit (Thermo Fisher Scientific) and stored at −20 °C. FFPE blocks were reviewed and annotated for tumour cellularity by a certified pathologist, with a minimum accepted tumour cellularity threshold of 20%. DNA concentrations were quantified using a Qubit 3 fluorometer using a dsDNA HS Assay kit (Thermo Fisher Scientific).

Libraries were prepared using a TruSight Oncology 500 (TSO500) assay (Illumina) according to the manufacturer's instructions. In brief, 100 ng DNA from each sample was sheared using a Covaris M220 ultrasonicator (Covaris). Genomic libraries were prepared using a Hamilton Microlab STAR system (Hamilton) and sequenced on an Illumina NextSeq550 instrument with a High Output kit v.2.5 (300 cycles).

Initial analysis was performed using DRAGEN TruSight Oncology 500 analysis software (v.2.6.2, Illumina).

Variants were annotated and classified according to AMP guidelines into tiers IA, IB, IIC, IID, III and IV[92]. Variants classified as tier IV, synonymous, non-coding or having no established biological impact were excluded using Velsera Clinical Genomics Workspace (CGW). Remaining variants were manually reviewed using in silico prediction tools (REVEL, CADD and SpliceAI) together with annotations from Franklin (Genoox), UCSC Genome Browser and Velsera CGW. Variants were reported only when sequencing coverage exceeded 100× and the variant allele frequency was >1%.

Epigenetic-focused CyTOF analysis of freshly isolated breast tumours, PDX models and breast cancer cell lines

CyTOF antibody panel. The antibody panel combined breast-tissue-relevant cell identity markers[33,34], histone PTMs spanning diverse chromatin states and the chromatin modifiers EZH2 and BMI-1 (Supplementary Table 1). The chromatin marks analysed were chosen on the basis of availability of highly specific and validated antibodies, and quantification of core histones was used for signal normalization.

Cell dissociation and CyTOF staining. Following surgery, a sample of the tumour tissue was kept in MACS Tissue storage solution (Miltenyi Biotec, 130-100-008), and dissociation into single cells was conducted using a human Tumour Dissociation kit (Miltenyi Biotec, 130-095-929) according to the manufacturer's protocol. Red blood cells were removed using Red Blood Cell Lysis solution (10x, Miltenyi Biotec, 130-094-183) and $CD45^+$ cells were depleted using human CD45 MicroBeads (Miltenyi Biotec, 130-045-801). The cell suspension was then washed with PBS and either further processed for CyTOF analysis or nuclei were extracted for snCUT&Tag analysis (when the number of cells was suitable). For cell lines, cells were dissociated using Accutase (Sigma-Aldrich, STC-07922) supplemented with DNase I (Merck, 4716728001).

For CyTOF analysis, cells were washed with Maxpar CyTOF PBS (Standard BioTools) and stained with 1.25 µM cisplatin (Standard BioTools) for 1 min. Cisplatin was neutralized with warm 10% FBS–DMEM. Cells were washed with Maxpar Cell Staining buffer (Standard BioTools) and blocked using TruStain FcX, Fc Receptor Blocking solution, human (422301, BioLegend), followed by 30 min of incubation with a cocktail of extracellular antibodies (Supplementary Table 1). For tumour samples, cells were washed with PBS, mildly fixed (1.6% paraformaldehyde (PFA)–PBS) for 30 min and frozen in 10% DMSO–FBS. When the number of cells allowed, aliquots of frozen cells from this step were used for additional applications (flow-cytometry-based sorting or WGBS). For cell lines, samples were directly stained for intracellular markers without mild fixation and freezing.

For intracellular staining and CyTOF measurements, tumour samples were thawed, all samples were washed with staining buffer, fixed and permeabilized using Maxpar nuclear antigen staining buffer set (Standard BioTools) according to the manufacturer's instructions. For multiplexed samples, barcoding was performed using a Cell-ID 20-Plex Pd Barcoding kit (Standard BioTools). In brief, each sample was resuspended in MaxPar Nuclear Antigen Staining Perm buffer (Standard BioTools) and mixed with freshly thawed barcode reagent. Samples were incubated for 60 min at room temperature with mixing every 15 min, washed twice with Perm buffer and all barcoded samples were combined into a single tube or well. Before antibody staining, cells were blocked using 10% FBS for 10 min and then incubated with the antibody cocktail (Supplementary Table 1) diluted in MaxPar Nuclear Antigen Perm buffer (Standard BioTools) for 30 min at room temperature. Cells were fixed with 4% PFA overnight. On acquisition day, samples were suspended in 4% PFA containing 125 nM Cell-ID Intercalator-Iridium (Standard BioTools), washed twice with staining buffer and twice with Maxpar Cell acquisition solution (201240, Standard BioTools). Samples were suspended in Maxpar Cell acquisition solution supplemented with 10% EQ Four Element Calibration Beads (201078, standard BioTools). Before acquisition on a CyTOF machine, cells were adjusted to about 300,000 cells per ml and filtered through a 35 µm mesh cell strainer. Data were acquired on a CyTOF Helios system (Standard BioTools).

The CyTOF software by Standard BioTools was used for normalization. Several gates were applied using FlowJo software: normalization beads were gated out using the 140Ce channel; live single cells were gated using the cisplatin 195Pt, iridium DNA label in 193Ir, event length and the Gaussian parameters of width, centre, offset and residual channels. De-barcoding was done using CyTOF software[93].

Epigenetic CyTOF data analysis

Pre-processing. Overall, 17 different tumour samples were analysed by CyTOF, with 5 samples analysed twice as either technical replicates in different batches to provide anchor information or with different antibody panels for the analysis of an additional cell-cycle marker (pRb). A total of 1,565,881 viable cells were measured (around 30,000–400,000 cells per sample; Supplementary Table 3). The remaining CD45+ cells were removed from the analysis. For PDX analysis, $CD45^+$ cells were removed and only human cells were analysed by maintaining human $CD298^+$ and mouse MHC class I-negative cells.

Batch correction. As samples were measured in batches, anchor samples, identical tumour samples measured in different batches, were used to implement batch-effect correction. Batch effects were corrected using a Q-Q transformation based on the empirical cumulative distribution function (ECDF). This method non-linearly maps quantile values from one anchor sample to corresponding quantiles in another, which effectively aligns their distributions. This transformation was then applied to tumour samples in the first batch.

Notably, as Q-Q plot transformation is inherently restricted to the range of the anchor samples owing to its reliance on interpolation, for values outside this range, linear extrapolation—based on the mean and standard deviation of the samples—can approximate the transformation's slope and intercept, which provides a reasonable estimate of the nonlinear mapping beyond the observed data range.

The transformation can be approximated by the following equation:

$$x' = \mathrm{ECDF}_2^{-1}(\mathrm{ECDF}_1(x)) \approx \beta_1 \times x + \beta_0 + \varepsilon \approx \left(\frac{\sigma_2}{\sigma_1}\right) \times x + \left(\mu_2 - \mu_1 \times \frac{\sigma_2}{\sigma_1}\right) + \varepsilon$$

Following batch correction, the full cohort was subsampled to roughly 10,000 cells per sample for subsequent analysis with stroma cells. Following the removal of stroma cells, the cohort was further subsampled to 5,000 cells per sample. Normalization and data scaling were performed on either all tumours as one batch or on individual tumours, as previously described[22,23]. In brief, cells were gated using the core histones (H3, H3.3 and H4), and the signals for intracellular proteins, including histone PTMs, was subsequently normalized using these core histone values to reduce systematic technical variation:

$$M^{ij} = \frac{M^{ij}}{(1 + \alpha\ \mathrm{H}^i_{3.3} + \beta\ \mathrm{H}^i_4 + \gamma \mathrm{H}^i_3)}$$

The coefficients were selected to minimize the sum of the variances of the core histones. UMAP representations show a random subset of data points (around 10%) to improve visualization.

Dimensionality reduction and clustering. To analyse the high-dimensional data generated from our CyTOF experiments, we first applied dimensionality reduction using UMAP[94]. This was followed by clustering with DBSCAN[95]. Clusters were then joined on the basis of similarity in expression patterns to define the three distinct cellular subpopulations: cycling, luminal and basal-like.

PermCell signature scoring using permutation-calibrated marker program scoring for single-cell data. We quantified, for each cell and each predefined gene signature, whether the set was enriched relative to a competitive null that matches the signature size but samples features from the same cell's remaining markers. The method returns per-cell *z*-scores, which were visualized on UMAP embedding.

Detailed computation is as follows:

Let $X \in R^{(N \times G)}$ be the cell-by-feature matrix. For sets $\mathcal{S} = \{S_1, \ldots, S_S\}$, each S_s has size m_s $m_s = |S_s|$. Optional signed feature weights $w_{\{s,g\}}$ are supported ($w_{\{s,g\}} = 0$ if $g \notin S_s$).

We row-centred per cell to remove global level shifts as follows:

$$X_c = X - 1\hat{\mu}^T, \hat{\mu}_i = \frac{1}{G}\sum_{g=1}^{G} X_{i,g}$$

Signed observed sum (unweighted) as follows:

$$\mathrm{Obs}_{i,s} = \sum_{g \in S_s} X_{c,i,g}$$

Signed observed sum (weighted) as follows:

$$\mathrm{Obs}_{i,s} = \sum_{g=1}^{G} w_{s,g} X_{c,i,g}$$

Absolute-magnitude observed sum (unweighted) as follows:

$$\mathrm{Obs}_{i,s}^{\mathrm{abs}} = \sum_{g \in S_s} |X_{c,i,g}|$$

Absolute-magnitude observed sum (weighted) as follows:

$$\mathrm{Obs}_{i,s}^{\mathrm{abs}} = \sum_{g=1}^{G} |w_{s,g}||X_{c,i,g}|$$

For competitive null (size-matched permutations or exact enumeration), for each set *s*, we formed a pool of competitor genes $\mathcal{P}_s$. By default we used the complement in the current feature space, $\mathcal{P}_s = \{1, \ldots, G\} \setminus S_s$ (exclude_set = True), enforcing a competitive null; if the complement was too small we fell back to all features. We then sampled m_s genes without replacement from $\mathcal{P}_s$ to construct null sums for each cell:

$$\mathrm{Null}_{i,s}^{(b),\mathrm{abs}} = \sum_{g \in \mathcal{B}_s^{(b)}} |X_{c,i,g}|$$

With $b = 1,\ldots,B$ permutations. When the number of combinations $C(|\mathcal{P}_s|, m_s)$ was below a threshold (default 5×10^4), we enumerated exactly in streamed chunks. For each cell and set, we estimated the null mean and variance (and analogues for the absolute variant) with numerically stable streaming updates; a small constant (10^{-6}) was added to standard deviations to avoid division by zero.

Per-cell *z*-scores (signed and absolute) were calculated as follows:

$$Z_{i,s} = \frac{\mathrm{obs}_{i,s} - \hat{\mu}_{i,s}}{\hat{\sigma}_{i,s}}, Z_{i,s}^{\mathrm{abs}} = \frac{\mathrm{obs}_{i,s}^{\mathrm{abs}} - \hat{\mu}_{i,s}^{\mathrm{abs}}}{\hat{\sigma}_{i,s}^{\mathrm{abs}}}$$

Right-tailed *P* values (two-sided optional) were calculated as follows:

$$P_{i,s} = 1 - \Phi(Z_{i,s}), P_{i,s}^{\mathrm{abs}} = 1 - \Phi(Z_{i,s}^{\mathrm{abs}})$$

To summarize effect size while preserving direction, we defined:

$$Z_{i,s}^{\mathrm{dir}} = \mathrm{sign}(\mathrm{obs}_{i,s}) \times Z_{i,s}^{\mathrm{abs}}$$

Thus, Z_dir is large and positive when the set is consistently up (large absolute signal with positive net direction), and large and negative when consistently down.

The framework supports signed feature weights $w_\{s,g\} \in \mathbb{R}$. Positive weights required higher expression, whereas negative weights encoded markers expected to be low in the target program. The null maintained the same size $m_s = \sum_g \mathbb{1}[w\{s,g\} \neq 0]$ by drawing m_s unweighted competitors from $\mathcal{P}_s$, preserving the competitive interpretation. Observed signed sums used $w_{\{s,g\}}$, whereas absolute-magnitude sums used $|w_{\{s,g\}}|$.

We used $B = 2{,}000$ permutations per set (10,000 for final figures), exclude_set = True (competitive null), two-sided *P* when depletion was relevant and FDR control at 5%. For scRNA-seq, we recommend computing on log-normalized counts after library-size normalization and optional regression of technical covariates; for CyTOF, arcsinh or *z*-score scaled channels work well.

Machine-learning classification of cells into subpopulations. In each tumour, a training set of cells was constructed, including only cells that were successfully classified as luminal, basal-like or cycling, keeping class-distribution and size equal between samples. From each tumour, 80% of these cells were randomly assigned to the training set, and the remaining 20% were used as the validation set. We used an XGBoost model[35], with a learning rate of 0.1, a maximum depth of 10 and 1,000 estimators. The model achieved a prediction accuracy of 94% on the validation set. The trained model was subsequently applied to cells that were not assigned to a subpopulation by DBSCAN (cluster −1), and cells predicted with >95% confidence were assigned to the corresponding luminal, basal-like or cycling subpopulation for downstream analyses (Extended Data Fig. 3b). The contribution of each feature to the XGBoost classification was evaluated using SHAP scores[30,31], which reflect feature contribution to the classification rather than relative expression levels.

Vendi score to determine epigenetic diversity. Vendi scores[96] were computed based on the epigenetic CyTOF panel (histone PTM and chromatin modifiers). For each sample, cells were stratified by class label. To control for differences in cell number across classes, we repeatedly subsampled $m = \lfloor n_\mathrm{min}/2 \rfloor$ cells without replacement from each class, where $n_$min was the smallest class size in that sample. Marker intensities were first normalized using a global *z*-score computed across a balanced sample of cells from all datasets. For each rarefaction, epigenetic marker values were discretized into 20 uniform bins, and the normalized dual Vendi score was computed on the resulting cell-by-marker matrix. Reported values correspond to the mean across 5,000 rarefactions, with uncertainty intervals defined by the 2.5th and 97.5th percentiles of the rarefied score distribution.

K5-based sorting followed by proteomic, transcriptomic, CUT&Tag and EM-seq analyses

Cell staining, sorting and sample preparation. Sample preparation for mass spectrometry analysis was optimized to enable intracellular staining before FACS with limited input material. Mildly fixed (1.6% PFA) breast tumour single-cell suspensions prepared for CyTOF were thawed and washed with Maxpar CyTOF PBS. Cells were fixed and permeabilized using Maxpar nuclear antigen staining buffer set (201063, Standard BioTools) according to the manufacturer's instructions. Staining was done in 10% FCS/CyTOF Nuclear Antigen Staining Perm with the following antibodies: K5–AF546 (Cell Signaling, 71536, in-house fluorophore conjugation) and pan-keratin–AF647 (Cell Signaling, 4528). Following three washes with PBS, cells were transferred to a new well, to minimize residual detergent content. Samples were read using FACS (BD Aria III), with 200 cells of each population ($K5^{hi}$ and $K5^{low}$; Supplementary Fig. 12a) sorted directly into freshly prepared 3 µl lysis buffer-containing wells in a 96-well plate format (Eppendorf twin.tec PCR plates). With a sorter droplet volume of 3.33 nl, the sorting process contributed an additional 0.66 µl to the total volume per well.

The cell lysis buffer was prepared by combining 5 mM tris-(2-carboxyethyl) phosphine (TCEP) as a reducing agent, 20 mM

chloroacetamide (CAA) alkylation agent, 100 mM triethylammonium bicarbonate (TEAB) and 0.1% *n*-dodecyl-β-D-maltoside (DDM). After sorting, the plate was briefly centrifuged to ensure cells settled into the lysis buffer. The plate was then sealed with PCR cap strips and incubated at 95 °C for 30 min in an Eppendorf ThermoMixer C, with the lid covered to prevent condensation. Following incubation, the plate was cooled, briefly centrifuged and 11 µl protein digestion buffer (50 mM ammonium bicarbonate buffer, 0.033 ng µl^{-1} sequencing grade modified trypsin (Promega) and 0.132 ng µl^{-1} Trypsin/Lys-C mix, mass spec grade (Promega)) was added to each well. The plate was resealed, briefly centrifuged and incubated overnight at room temperature for protein digestion. The next day, 20 µl solvent A (99.9% water and 0.1% formic acid) was added to each well to acidify the samples and halt the trypsinization process.

LC–MS analysis. LC–MS analyses were performed on an Evosep1 liquid chromatography system (Evosep) coupled to the TimsTOF SCP mass spectrometer (Bruker). The samples were loaded onto a disposable Evotip C18 trap column (Evosep Biosystems) according to the manufacturer's instructions. In brief, the workflow involved rinsing dry Evotips with 20 µl solvent B (99.9% acetonitrile and 0.1% formic acid), followed by centrifugation at 800*g* for 60 s. The tips were then conditioned by soaking in 100 µl 1-propanol until they appeared pale white. After conditioning, the tips were equilibrated with 20 µl solvent A, followed by centrifugation at 800*g* for 60 s. The protein samples were loaded onto the wet tips, followed by another centrifugation step at 800*g* for 60 s. To ensure cleanliness, the tips were washed with 20 µl solvent A, with centrifugation repeated. Finally, to maintain tip hydration, 100 µl solvent A was added, and the tips were centrifuged briefly at 800*g* for 10 s.

Peptides were separated on an Aurora Elite CSI 15 × 75 C18 UHPLC column (IonOpticks) using the Whisper 40 SPD method. Mass spectrometric acquisition was performed in a data-independent (DIA) PASEF mode. For the diaPASEF mode, the ion accumulation and ramp time was set to 100 ms, and precursors in an ion mobility range from 1/K0 = 1.45 V s^{-1} cm^{-2} to 0.64 V s^{-1} cm^{-2} were analysed over a mass to charge (*m/z*) range from 100 to 1,700. Precursor ions for MS/MS analysis were isolated based on diaPASEF window placement covering an ion mobility range from 1/K0 = 0.75 V s^{-1} cm^{-2} to 1.2 V s^{-1} cm^{-2} and a *m/z* range of 400–1,000. The estimated cycle time was 0.64 s with 1 MS ramp and 5 MS/MS ramps, resulting in 75 MS/MS windows. The collision energy was ramped from 20 eV at 1/K0 = 0.6 V s^{-1} cm^{-2} to 59 eV at 1/K0 = 1.6 V s^{-1} cm^{-2}.

Data analysis. DIA-NN (v.1.8.1) was used to analyse the mass spectrometric raw files[97]. Analyses of the files were performed using a library-based approach (spectral library provided by BRUKER containing 67,686 proteins and 13,827 genes). Library precursors were re-annotated using a FASTA database (FASTA files UP000000589_10090.fasta and UP000000589_10090_additional.fasta, version August 2022). Trypsin was selected as protease, with a maximum of one missed cleavage. DIA-NN default settings were used with the following exceptions: the charge range of precursors was set to 2–10, and the 'unrelated runs' option was selected. A total of 5,630 proteins were identified.

Functional enrichment analysis of differentially expressed proteins was done using the enricher function (clusterProfiler package[98]) comparing to MsigDB hallmark gene sets[85].

H3K4me1 CUT&Tag on sorted populations. Single-cell suspensions of mildly fixed (1.6% PFA) cells derived from luminal PDX tumours or MCF7 cultures were stained as described above. For PDX samples, cells were stained with K5–AF555 (Cell Signaling, 72825) and pan-keratin–AF647 (Cell Signaling, 4528). Pan-keratin-positive epithelial cells were sorted on the basis of K5 expression levels into K5^{hi} and K5^{low} populations. For MCF7 cells, staining was performed using K5–AF555 and GATA3–Brilliant Violet (BioLegend, 653814). Cells expressing high K5 and low GATA3 were defined as K5^{hi} and compared with the remaining cell population (K5^{low}). Cells were sorted using a Sony SH800S cell sorter (Sony Biotechnology) with single-stain controls for compensation and to establish gating thresholds.

Following sorting, H3K4me1 profiling was performed using a CUT&Tag-IT Express Assay kit (Active Motif, 53175) according to the manufacturer's instructions. The H3K4me1 antibody (Cell Signaling, 5326) was used at a 1:50 dilution. Decrosslinking was performed by incubation for 1 h at 55 °C followed by 4 h at 65 °C. Libraries were sequenced on an AVITI platform (Element Biosciences) using paired-end 50 bp reads. Sequencing data were demultiplexed and aligned to the hg38 assembly. Peaks were called using MACS2 with the --keep-dup 1 parameter. Peaks unique to either K5^{hi} or K5^{low} samples were assigned to genes using the ChIPseeker R package. Functional enrichment analysis was performed using the enricher function from the clusterProfiler R package, using gene signatures detailed in Supplementary Table 4.

EM-seq of sorted populations. EM-seq was adapted for low-input samples. Tumour single-cell suspensions were stained with K5 and pan-keratin antibodies as described above. In brief, 200 cells of each population (K5^{hi} and K5^{low}) were sorted directly into freshly prepared lysis buffer (10 mM Tris-HCl, pH 8.0, 200 mM NaCl, 10 mM EDTA, 0.5% SDS and 0.3 mg ml^{-1} proteinase K) and frozen. For each condition, four replicates were analysed.

Lysed cells were incubated at 55 °C for 3 h, followed by overnight incubation at 60 °C to reverse crosslinking. Tris-HCl and NaCl were added to reduce the SDS concentration to 0.33% before PCI extraction. The aqueous phase was recovered, NaCl was adjusted to 300 mM and glycogen added to 0.13 mg ml^{-1} following DNA precipitation with 2.3 volumes of ice-cold ethanol at −20 °C overnight. DNA was pelleted by centrifugation (21,000*g*, 45 min, 4 °C), washed once with 70% ethanol and resuspended in 10 mM Tris-HCl, 0.1 mM EDTA (pH 8.0).

DNA samples were enzymatically fragmented using a NEBNext UltraShear kit (M7634S, New England Biolabs) for 25 min at 37 °C. Libraries were prepared using a NEBNext Enzymatic Methyl-seq v.2 kit (E8015S, New England Biolabs) and NEBNext LV Unique Dual Index Primers Set 2A (E3390S, New England Biolabs) according to section 3 of the NEBNext UltraShear user manual, with the following modifications: unmethylated lambda and CpG methylated pUC19 DNA controls were diluted 1:500. Libraries for replicates three out of the four replicates were prepared with conversion controls. DNA denaturation before cytosine deamination was performed with formamide (4401457, Thermo Fisher Scientific). Libraries were amplified for 12 PCR cycles and purified by 2 rounds of bead cleanup with NEBNext Sample Purification Beads at 0.8× ratio. Library quality and fragment size distribution were assessed on an Agilent TapeStation with a High Sensitivity D5000 kit, which produced an average fragment size of about 500 bp. Libraries were sequenced on AVITI high-output flow cells (Element Biosciences), which generated 150-bp paired-end reads.

Paired-end sequencing reads were subjected to adapter and quality trimming using cutadapt[99] (v.4.4; parameters: --quality-cutoff 20 --overlap 5 --minimum-length 25; Illumina TruSeq adapter 'AGATCGGAAGAGC' clipped from both reads), followed by trimming of 10 and 5 nucleotides from the 5′ and 3′ end of the first read and 15 and 5 nucleotides from the 5′ and 3′ end of the second read. The trimmed reads were aligned to the human reference genome (hg19) using BSMAP[100] (v.2.90; parameters: -v 0.1 -s 16 -q 20 -w 100 -S 1 -u -R). Duplicates were removed using the 'MarkDuplicates' command from GATK[101] (v.4.6.2.0; --VALIDATION_STRINGENCY = LENIENT --REMOVE_DUPLICATES=true). Methylation rates were called using mcall from the MOABS package[102] (v.1.3.9.6; default parameters). To assess similarity between sorted tumour populations, pairwise correlations of CpG methylation levels were calculated using only shared CpG present in all eight samples of a given tumour. Pearson's correlation coefficients were calculated across these shared CpG sites.

Differential methylation analysis comparing $K5^{hi}$ and $K5^{low}$ cells was conducted for each tumour separately, with replicates pooled to obtain higher coverage. WGBStools (https://doi.org/10.26508/lsa.202503514) was used to identify differentially methylated regions, which were assigned to genes based on promoter overlap (−2,000 to +500 from the transcription start site (TSS)). For each gene, promoter-associated methylation changes were aggregated, and only genes with an absolute difference >0.2 and FDR-corrected Fisher's exact test <0.05 were considered. Genes were classified according to the predominant direction of methylation change.

Functional enrichment analysis was performed using the enricher function (clusterProfiler package), comparing to MsigDB hallmark gene sets or curated breast cancer cell identity signatures.

Replicate-pooled EM-seq methylation profiles were summarized at the CGI level for each tumour and state combination by calculating CGI methylation as the ratio of methylated reads to total coverage. TCGA BRCA DNA-methylation data (Illumina HumanMethylation450) were mapped to hg19 CGI coordinates, and CGI methylation for each tumour was defined as the mean β value across probes overlapping the corresponding CGI. A shared CGI feature space was constructed as the intersection between CGIs detected in EM-seq and those measurable in TCGA. In each sample (tumour 5 or tumour 23), differential methylation between $K5^{hi}$ and $K5^{low}$ states was calculated for each CGI, and CGIs showing concordant direction of change across the two samples were retained. The effect size for each retained CGI was defined as the median of the model-specific differences and used to assign a signed weight. TCGA CGI methylation values were standardized across tumours, and for each tumour, a signature score was computed as the weighted mean of standardized CGI methylation values across the retained CGIs. Signature scores were compared across PAM50 subtypes.

Read-level DNA methylation heterogeneity was quantified using the RLM tool[103]. For each replicate, BSMAP-aligned paired-end BAM files were analysed against the hg19 reference genome using parameters -m PE, -a bsmap, and -s all. Only loci with sequencing coverage ≥10 reads were retained for analysis. Entropy values from all loci were aggregated for each sample and compared between the $K5^{hi}$ and $K5^{low}$ in each tumour sample.

RNA extraction and bulk MARS-seq

For bulk MARS-seq analysis, MCF7 cells were sorted into $K5^{hi}GATA3^{low}$ and $K5^{low}GATA3^{hi}$ populations (as described above) to capture basal-like and extreme luminal subpopulations. Cells were dissociated using Accutase (Sigma-Aldrich, STC-07922) supplemented with DNase I (Merck, 4716728001), washed with PBS and viability assessment was conducted using Zombie Violet fixable viability dye (BioLegend, 423105 or 423113). Cells were incubated for 10 min on ice in the dark, followed by 2 washes with staining buffer (PBS containing 2% FBS and 2 mM EDTA). For intracellular staining, cells were fixed in 4% formaldehyde–PBS for 15 min at room temperature. The fixation reaction was quenched with glycine (final concentration 125 mM), and cells were centrifuged at 700*g* for 5 min at 4 °C, then washed once with ice-cold staining buffer and once with PBS. Cells were permeabilized with 0.1% Triton X-100 in PBS for 15 min at room temperature, washed twice with staining buffer, and blocked for 10 min on ice with blocking buffer (PBS containing 5% FBS and 2 mM EDTA). Cells were then stained with anti-GATA3-BV421 (1:10 dilution, BLG-653814) and anti-KRT5-Alexa Fluor 555 (1:20 dilution, CST-72825S) antibodies in staining buffer overnight at 4 °C protected from light, followed by two washes with sorting buffer. Flow cytometry was performed on a Sony SH800S Cell Sorter. Single viable cells were gated (Zombie dye exclusion). Intracellular GATA3 and KRT5 expression were quantified in viable cell populations using single-stain controls for compensation and to establish gating thresholds (Supplementary Fig. 12b).

Total RNA was extracted using a RNeasy Micro Kit (Qiagen, 74004) according to the manufacturer's instructions and stored at −80 °C. Bulk MARS-seq libraries were prepared from extracted RNA following the published protocol[104] and sequenced on a NextSeq 500 instrument using a NextSeq 500/550 High Output Kit v.2.5 (75 cycles). Differential expression analysis was conducted using UTAP transcriptome analysis pipeline[105].

Flow cytometry sorting of $CD74^{hi}$ and $CD74^{low}$ MCF7 cells. MCF7 cells were dissociated using Accutase (Sigma-Aldrich, STC-07922) supplemented with DNase I (Merck, 4716728001) and counted in culture medium. Cells were washed with PBS and viability assessment was conducted using Zombie NIR or Zombie Violet fixable viability dye (BioLegend, 423105 or 423113). Cells were incubated for 10 min on ice in the dark, followed by two washes with staining buffer (PBS containing 2% FBS and 2 mM EDTA). For surface marker staining, cells were incubated with anti-human CD74–FITC antibody (1:10 dilution in staining buffer; eBioscience, 11-0748-42, clone 5-329) for 45 min at 4 °C in the dark with intermittent mixing every 10 min, then washed twice with sorting buffer (PBS containing 1% FBS and 5 mM EDTA). Viable cells were sorted into collection medium containing 50% FBS and 50% DMEM supplemented with 1% penicillin–streptomycin. To enrich for cells of the basal-like population, the top 10% of cells with the highest CD74 expression ($CD74^{hi}$) and, as a control, the bottom 10% of cells with the lowest CD74 expression ($CD74^{low}$) were sorted based on CD74–FITC fluorescence intensity (Supplementary Fig. 12c). Cells were sorted using a Sony SH800S cell sorter (Sony Biotechnology) with single-stain controls for compensation and to establish gating thresholds.

Migration assay

Transwell assay. Cell migration capacity was assessed using Transwell polycarbonate membrane inserts (6.5 mm diameter, 8.0 µm pore size; Corning, 3422) placed in 24-well plates. Sorted cells were seeded in complete culture medium and allowed to recover for approximately 20 h. Cells were then serum-starved in DMEM supplemented with 0.1% FBS for 5 h before the assay. Following serum starvation, cells were dissociated using Accutase, resuspended in DMEM with 0.1% FBS and counted. A total of 5×10^4 cells in 150 µl DMEM with 0.1% FBS were seeded into the upper chamber. The lower chamber was filled with 600 µl of medium containing 20% FBS and 100 ng ml^{-1} epidermal growth factor (EGF; PeproTech, AF-100-15) as a chemoattractant. Control wells contained DMEM with 0.1% FBS in both chambers to assess random migration.

Following a 20-h migration period, non-migrated cells on the top surface of the membrane were removed with a cotton swab, and migrated cells on the bottom surface were fixed with 4% formaldehyde (Thermo Fisher Scientific, 28906) for 15 min at room temperature and stained with 0.1% crystal violet (Sigma-Aldrich, C0775) in 20% methanol for 10 min. Membranes were washed with distilled water and air-dried. Owing to weak crystal violet signal intensity and high background noise that precluded accurate automated quantification, migrated cells were imaged using the Cy5 channel (excitation and emission of 640 and 670 nm, respectively) to detect crystal violet autofluorescence. This approach enabled reliable background subtraction and automated cell counting. The Cy5 signal showed complete overlap with crystal-violet-positive cells, thereby confirming that autofluorescence accurately represented the migrated cell population. Quantification was performed using ImageJ software (v.1.53; National Institutes of Health), and results are expressed as the number of migrated cells per field.

Wound-healing assay. Collective cell migration and wound closure were assessed using a two-dimensional wound-healing assay. Culture-Inserts 2 Well for self-insertion (ibidi, 80209) were placed in 6-well plates (Corning, CC-3516) under sterile conditions. Cells were seeded at 1×10^5 cells per well in 70 µl of complete growth medium and incubated for 24 h until a confluent monolayer formed. Culture inserts were carefully removed with sterile tweezers to create a defined cell-free gap of approximately 500 µm width. Wells were washed twice

with PBS to remove non-adherent cells, and 2 ml of fresh medium was added. Images were acquired at 0 h, 20 h, 46 h and 54 h after wounding at ×10 magnification using an inverted microscope.

Wound areas were quantified from bright-field images using an interactive segmentation workflow based on the Segment Anything Model (SAM). For each image, the wound region was identified as the continuous cell-free area and segmented using user-guided point prompts placed in the wound and, when required, outside the wound to suppress surrounding tissue or background. The resulting binary mask was visually inspected and refined interactively until it matched the wound boundary. The wound area was then calculated as the number of pixels contained in the final mask. To account for differences in image size or field of view, the fractional wound area was computed for each image by dividing the wound-mask area by the total image area. When spatial calibration was available, pixel areas were converted to physical units (μm^2) using the microscope pixel size. This procedure yielded both absolute and normalized wound-area measurements for comparison across experimental conditions.

WGBS of breast tumour samples

Sample processing, bisulfite reaction and sequencing. Mildly fixed breast tumour single-cell suspensions prepared for CyTOF were thawed and DNA was extracted using a NucleoSpin Tissue kit. RNA was removed by the addition of 20 µl of 10 mg ml^{-1} RNase during the lysis step. The generation of dual-indexed libraries WGBS followed a modified protocol based on the NEBNext Ultra DNA Library Prep kit for Illumina (New England BioLabs, E7442, E7445 and E7535), in accordance with the manufacturer's instructions. The library input was 500 ng gDNA, quantified using a Qubit BR assay (Invitrogen, Q32853), and 1% unmethylated Lambda DNA (Promega, D1521) was added as a control for bisulfite conversion efficiency. Fragmentation of input gDNA to achieve an average insert size of 350 bp was done with a Covaris S220 Focused-ultrasonicator. Size selection was performed using SPRIselect beads, isolating fragments in the 300–400 bp range. Bisulfite conversion after size selection was done using an EZ DNA Methylation-Lightning kit (Zymo, D5030), following the manufacturer's guidelines. For amplification after bisulfite conversion, Kapa HiFi Uracil+ (Kapa Biosystems, KK282) polymerase was used under the following cycling conditions: 98 °C for 45 s, followed by 8 cycles of 98 °C for 15 s, 65 °C for 30 s, and 72 °C for 30 s, with a final extension step at 72 °C for 1 min. WGBS libraries were sequenced on an Illumina NovaSeq 6000 instrument using an S4 flowcell and a 150 bp paired-end dual indexed run, with 5% PhiX library control (Illumina). The resulting coverage ranged from 16.7× to 22.5×. The bisulfite conversion rate for unmethylated Lambda DNA averaged 99.8%.

WGBS data analysis. Paired-end sequencing reads were pre-processed with Trim Galore! (v.0.6.5) before alignment with Bismark (v.0.24.2) and Bowtie2 (v.2.3.5.1)[106] using the 'non_directional' option. A reference genome was generated for bisulfite-converted human hg19 (obtained from UCSC). The resulting alignment files were merged and sorted using Samtools (v.1.16.1) and further deduplicated with 'deduplicate_bismark', before calling of context-dependent methylation with 'bismark_methylation_extractor'.

Empirical entropy was calculated from read-level methylation patterns using ClubCpG package[55]. Read clustering of methylation patterns over 100 bp genomic bins was used to determine the observed empirical distribution of DNA methylation states, without imputation for reads that did not fully cover a given bin. As there is no reference methylation pattern for individual cell types in this dataset, we used an empirical entropy estimate to gauge the degree of cell type heterogeneity in each sample, using as the individual pattern probabilities the proportion of reads with that state, and normalizing $P\log(P)$ over 2^n, where n is the number of CpGs in the bin. Each bin was assigned a normalized entropy score. Genomic bins were associated with genes (±3 kb from the TSS) using ChIPseeker package[107]. A list of CGIs was downloaded from the UCSC table browser[108] and bedtools intersect function was used to associate with genes (−1,000 bp from TSS − TES (transcription end site), n = 28,170)[109]. The average entropy was calculated for each CGI, and the functional enrichment for the associated genes was calculated using the enricher function (clusterProfiler package[98]) comparing with MsigDB genes sets[85].

To deconvolute the DNA methylation data compared with the TCGA BRCA dataset[51], the methylation status was calculated for each probe in the HumanMethylation450 array (Illumina) using the HumanMethylation450 v.1.2 Manifest File (https://emea.support.illumina.com/downloads/humanmethylation450_15017482_v1-2_product_files.html). The combined dataset (TCGA and this study) was then deconvoluted to different numbers of archetypes using the EDec package (https://github.com/BRL-BCM/EDec)[110]. The PAM50 information in the TCGA BRCA dataset was used to annotate tumours by subtype. The associated expression data available for the TCGA BRCA dataset were used to correlate expression patterns with each archetype. Genes were ranked according to the Pearson's correlation coefficient for each archetype, which enabled us to associate each archetype with functional terms using GSEA and the MSigDB Hallmarks gene sets[85,111,112].

Analysis of single-cell multiome datasets

Simultaneous profiling of chromatin accessibility and gene expression was performed using the Chromium Next GEM Single Cell Multiome ATAC + Gene Expression platform (10x Genomics) following the manufacturer's protocol. Cryopreserved breast cancer cells were rapidly thawed, gradually diluted in RPMI supplemented with 10% FBS, washed in PBS containing 0.04% BSA and filtered through a 40 µm strainer. Nuclei were isolated using the 10x Genomics nuclei isolation workflow, washed in chilled nucleus wash buffer and resuspended in diluted nucleus buffer. Nucleus suspensions were filtered to remove aggregates and adjusted to the recommended concentration before loading. Approximately 500 nuclei per sample were loaded onto a Chromium controller for tagmentation, barcoding and GEM generation. Gene expression and ATAC libraries were generated according to the manufacturer's protocol and indexed using 10x Genomics dual-index plates. Library quality and concentration were assessed using Qubit fluorometry and an Agilent TapeStation. Sequencing was performed on an Aviti platform according to 10x Genomics specifications. ATAC–seq data were processed with cellranger-arc (v.2.2.0) and mapped against human genome GRCh38. Gene expression data were processed using cellranger (v.10.0.0).

Previously published scRNA-seq and scATAC–seq data[13] were analysed as described by the authors. Only cells annotated as cancer cells, as defined by RNA profiles, were included in the analysis. To visualize gene expression pattern, UCell scores for the indicated signatures were calculated. To account for dropout, the gene signatures were smoothed using a Gaussian kernel with a bandwidth set to the 5th percentile of the Euclidean distance of the point-to-point distances in the normalized gene space. For the basal–luminal index, the z-score difference between the down in luminal versus basal and the up in luminal versus basal signatures[59] was computed. For the ATAC–seq data, peaks were called for each sample (only cancer cells) separately, and the WIP fraction was calculated. A lower WIP fraction, corresponding to a greater proportion of fragments outside called peaks, was interpreted as increased epigenetic noise[58].

snCUT&Tag profiling of H3K4 and H3K9 methylation in breast tumours

Sample preparation. Following CD45 depletion, 1×10^6 cells were resuspended in antibody buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 2 mM EDTA, 0.5 mM spermidine, 0.05% digitonin, 0.01% NP-40, protease inhibitors and 1% BSA), incubated on ice for 3 min and centrifuged for 3 min at 600g to extract nuclei as previously described[113]. Nuclei were

resuspended in antibody buffer supplemented with 10% DMSO and cryopreserved for further analysis.

For CUT&Tag, all washes were made with 500 µl unless otherwise stated and all centrifugations were done using a swinging bucket centrifuge at 600*g* for 8 min at 4 °C. Around 50,000 to 100,000 nuclei were resuspended in 50 µl antibody buffer (EDTA 2 mM, 20 mM HEPES pH 7.2, NaCl 150 mM, spermidine 0.5 mM, BSA 2%, NP-40 0.01%, digitonin 0.01% and proteases inhibitor 1×) with 1:50 antibody dilution (anti-H3K4me1 D1A9, 5326 and anti-H3K4me3 C42D8, 9751, Cell Signaling) or 0.3 µg (anti-H3K9me2 D85B4 80806, Cell Signaling) and incubated overnight at 4 °C with rotation. Nuclei were washed the next day with Dig-Wash buffer and resuspended in 100 µl Dig-300 buffer (20 mM HEPES pH 7.2, 300 mM NaCl, 0.5 mM spermidine, 2% BSA, 0.01% NP-40 and 0.01% digitonin) together with protein A–Tn5 fusion (Diagenode, C01070001, 1:250) and incubated for 1 h at room temperature with rotation. Nuclei were washed 3 times with Dig-300 buffer, resuspended in 300 µl Tag-Buffer (20 mM HEPES pH 7.2, 300 mM NaCl, 0.5 mM spermidine, 2% BSA, 0.01% NP-40, 0.01% digitonin and 10 mM $MgCl_2$) and incubated for 1 h at 37 °C. Tagmentation was stopped by addition of one volume of 1x diluted nuclei buffer (DNB, 10x Genomics) supplemented with 2% BSA and 12.5 mM EDTA. The nuclei were then centrifuged at 1,300*g* for 4 min at 4 °C and washed twice with 200 µl 1× DNB supplemented with 2% BSA. The nuclei were resuspended in 10–70 µl of DNB + 2% BSA. Nuclei were loaded on a 10x Chromium system using a 10X Chromium Single Cell ATAC–Seq kit v.2 (10x Genomics) as previously described[113]. GEM incubation, barcoding and amplification was performed according to the Chromium Single Cell ATAC Library kit manual with 17 PCR cycles. Libraries were amplified with 13 PCR cycles. snCUT&Tag libraries were sequenced on a NovaSeq 6000 (Illumina) in PE50 mode.

snCUT&Tag data analysis. Raw sequencing files (Fastq files) were processed using an in-house pipeline (https://github.com/vallotlab/scCUT-Tag_10X) to generate fragment files and to count matrices. In brief, the barcodes are aligned to indexed reference libraries using Bowtie2 (v.2.5.4) and the genomic parts of the reads are aligned to the human reference genome (hg38) using STAR (v.2.7.8a). Duplicate reads, reads mapping to the ENCODE blacklisted regions and reads with low mapping quality (q20) were filtered out.

Next, 50-kb bin matrices were used to generate a joint UMAP representation of all three tumours samples. Tumour microenvironment and tumour cells were annotated based on H3K4me1 or H3K4me3 enrichment at known marker genes (Supplementary Table 7). Owing to the physical removal of $CD45^+$ cells during tumour dissociation, no immune cells could be identified based on H3K4me1 or H3K4me3 profiles. Peaks were called using CallPeaks (Signac)[114]. Gene activity matrices were generated by aggregating counts across gene bodies and promoter regions extending (−2 kb from the TSS), following the standard Signac approach. These gene-level activity scores were used as features for dimensionality reduction and clustering. Count datasets were normalized using SCTransform[115] for each tumour separately. Differential binding occupancies were calculated by two-sided Wilcoxon signed-rank test (FDR-corrected). Functional enrichment was conducted using the enricher function (clusterProfiler package[98]), with MsigDB gene sets[85,116] as references. Epigenomic gene signature scores were computed using UCell[117] for each gene set of interest, starting from SCTransform counts. For visualization of gene signatures in UMAP space, to account for dropout, epigenomic gene signature scores were smoothed using a Gaussian kernel with a bandwidth set to the 5th percentile of the Euclidean distance of the point to point distances in the normalized gene space.

To identify loci that contribute to epigenomic intratumour heterogeneity (either H3K4me3 or H3K4me1), the SHAP algorithm was used on a trained XGBoost classifier[35]. For each tumour, the most contributing genes that led to accurate classification (around 400 genes, about 90% accuracy) were compared against gene signatures from MSigDB[85,116].

Pseudo-bulk coverage tracks were generated by aggregating fragment-level snCUT&Tag data from cells assigned to each cluster. For each sample and histone mark, fragments corresponding to the basal-like cluster and to all remaining clusters were pooled separately. To enable direct comparison, the non-basal-like group was downsampled to match the total fragment count of the basal-like population. Genome-wide coverage was computed from fragment intervals and normalized to counts per million (CPM), producing bedGraph files that were subsequently converted to bigWig format for visualization. Tracks were visualized using pyGenomeTracks.

Pseudobulk chromatin profiles were generated from single-cell data and used for peak calling with MACS3 (ref. 118), with narrow peak mode applied to H3K4me3 and broad peak mode to H3K4me1. Unified peak sets for each sample and histone modification were annotated using ChIPseeker[107], and genomic feature distributions were summarized as percentages of total peaks per sample, with promoter subcategories merged before visualization. Differential association of peaks to clusters was evaluated in a one-versus-rest framework. For each peak, signal presence was binarized and detection fractions were compared between the target population (cluster) and remaining cells using a one-sided two-proportion *z*-test. Effect sizes were computed as $\log_2$ fold change of CPM-normalized mean signal. *P* values were adjusted using the Benjamini–Hochberg procedure, and peaks were considered significant at adjusted $P \leq 0.05$, $\log_2$ fold change ≥ 0.25, and a minimum detection rate of 1%.

Motif enrichment analysis was performed using the Signac framework[114] with transcription factor binding profiles obtained from the JASPAR database[119]. Motifs associated with a pre-defined list of EMT and basal-like state regulators, defined separately for enhancer (H3K4me1) and promoter (H3K4me3) signals, were examined, using a nominal significance threshold of $P < 0.05$. Multiple testing correction was computed using the Benjamini–Hochberg procedure in the subset of tested transcription factors per sample, which reflect the reduced hypothesis space.

To compare the read count with gene activity, for each tumour the number of counts per cell ('nCount_RNA') per cell was correlated (Pearson's) with the epigenomic gene signature score (UCell computed on TFIDF normalized[115] data).

Using scGLUE[120], we performed multimodal integration of the H3K4me1 snCUT&Tag dataset with the CyTOF dataset. For each of the cell-identity markers present in the CyTOF antibody panel, we identified the relevant gene in the gene activity matrix and used those to form the weakly informative connection for scGLUE. After running scGLUE, we trained an XGBoost classifier on the CyTOF data in the joint latent space to predict the cell label and used the classifier to perform label transfer to the snCUT&Tag modality.

For cell annotation of the H3K9me2 dataset, the complementary H3K4me1 dataset from the same nucleus extraction was used. In brief, 20-kb bin matrices were generated for both H3K4me1 and H3K9me2 datasets, keeping cell labels derived from the H3K4me1 analysis. A classifier was trained on the labelled H3K4me1 dataset, which achieved >95% accuracy, and applied to the inverted *z*-score matrix of H3K9me2, using the generally opposing genomic distributions of H3K4me1-associated regulatory chromatin and H3K9me2-associated repressive chromatin as an inference framework[62,63].

For gene association analysis, H3K9me2 was analysed without peak calling, as this mark forms broad repressive domains rather than sharp focal peaks. Instead, promoter regions (−2 kb to +200 bp relative to the TSS) were tiled into 200 bp windows, and fragment counts were quantified across these promoter-proximal tiles to generate a promoter occupancy matrix. The resulting matrix was normalized using SCTransform[115] and used for UMAP embedding and clustering. Because H3K9me2 is a repressive histone mark, high gene-set scores were interpreted as increased repressive occupancy over the corresponding loci rather than increased transcriptional activity.

Plasma collection and EPINUC analysis

Blood samples were collected in EDTA tubes and transferred ambient for processing. Plasma was separated by centrifugation and stored at −80 °C.

Surface preparation for single-molecule imaging. PEGylated-biotin-coated and PEGylated-poly(T)-coated microscope slides were prepared as previously described[28,29]. In brief, glass coverslips were cleaned extensively (sonication rounds with acetone and Alconox), hydroxylated by incubation in 1 M KOH and passivated with amino-silanization chemistry (3% 3-aminopropyltriethoxysilane, 5% acetic acid in HPLC ethanol). Surface functionalization along with the first passivation step was performed by applying mPEG: PEGylated-biotin/PEG-azide solution (10:1 ratio). PEGylated surfaces were further passivated with MS(PEG)4. Anchoring of dT50 to PEGylated-poly(T) surfaces was done via click chemistry (10 μM 5′heyxynyl-dT50 (IDT), 2 mM $CuSO_4$).

EPINUC antibody labelling. Antibodies were labelled using a biotin conjugation kit (Abcam, ab201796) and Alexa Flour antibody-labelling kits (Thermo Fisher Scientific, A20181, A20183 and A20186) according to the manufacturer's protocol. Labelled antibodies were further purified by size-exclusion chromatography using Bio-Spin 6 columns (Bio-Rad, 7326200). Antibodies are listed in Supplementary Table 5a.

Single-molecule imaging of biomarkers. Quantification of the five protein biomarkers was conducted as previously described[28]. PEGylated-poly(T)-coated coverslips were assembled into a ChipSop flowcell (chip 2-28 mini Luer platform Fl. 342, 30001180), which generated a 14-lane flowcell. Streptavidin solution (0.2 mg ml^{-1} in PBS) was added to the surface for 15 min and washed with IMB2 (10 mM MES pH 6.5, 60 mM KCl, 0.32 mM EDTA, 3 mM $MgCl_2$, 10% glycerol, 0.1 mg ml^{-1} BSA and 0.02% Igepal). Biotinylated antibodies were incubated on the surface in IMB2 for 30 min, followed by washing with IMB2. Next, the plasma sample was added to the flowcell and incubated on the surface for 30 min, followed by three washes with IMB2 to allow binding of target proteins. Fluorescently labelled antibodies (detection antibodies) were introduced to the surface for 60 min, washed with IMB2 and imaged.

Imaging was performed on a TIRF microscope manufactured by Nikon (Eclipse Ti2-E LU-N4 TIRF) with a CFI apochromat TIRF ×60 objective lens and TRF49904, TRF49909 and TRF89902 filter cubes (Chroma) for the 488 mm, 561 mm and 647 nm lasers, respectively. Images were taken with the ×1.5 magnification setting, which resulted in fields of view (FOVs) of 148 μm × 148 μm using an Andor Zyla 4.2 Plus camera. At least 50 FOVs were imaged per lane.

Single-molecule imaging of cell-free nucleosomes histone PTMs. PEGylated-poly(T)-coated coverslips were assembled into an Ibidi flowcell (Sticky Slide VI hydrophobic, IBIDI), generating a 6-lane flowcell and further passivated with 5% BSA for at least 30 min.

Plasma nucleosomes were processed as previously described[28] with minor modifications. Nucleosomes were polyA tailed by the following reaction: 20 μl plasma was mixed with 10 μl of 10× Green buffer (Enzymatics), 416 μM $CoCl_2$ (Enzymatics), 1:60 protease inhibitor cocktail (Sigma), 83.3 μM dATP (Thermo Fisher Scientific), 5 μl Klenow Fragment (3′→5′ exo-, NEB), 3 μl T4 polynucleotide kinase (NEB) and 4 μl terminal deoxynucleotidyl transferase (TdT, Enzymatics). Following incubation at 37 °C for 1 h, TdT in the samples was inactivated by the addition of EDTA at final concentration of 5 mM.

Next, plasma samples containing polyA-tailed cell-free nucleosomes were incubated with antibodies (final concentration 1:60) for 30 min at room temperature to allow for the formation of antibody–cell-free nucleosome complexes. Next, samples were loaded on the surface and incubated for 15 min to allow hybridization of polyA tails to dT50 oligonucleotides. The flowcell was washed once with IMB (12 mM HEPES pH 8 (Thermo Fisher Scientific), 40 mM Tris pH 7.5 (Gibco), 60 mM KCl (Sigma), 0.32 mM EDTA (Invitrogen), 3 mM MgCl2 (Sigma), 10% glycerol (Bio-Lab), 0.1 mg ml^{-1} BSA (Sigma) and 0.02% Igepal (Sigma)), and cell-free nucleosomes were visualized by hybridizing 25 pM Cy3-dT23 in IMB to the free portion of polyA tails for 10 min. We supplemented Cy3-dT23 solution with TetraSpeck beads (Thermo Fisher Scientific), used as fiducial markers to allow correction for stage drift in image analysis. The flowcell was washed once with IMB, then filled with IMB, followed by time-lapse imaging of 5 cycles every 15 min with the 3 laser channels across all positions.

Imaging was performed on a TIRF microscope as described above obtaining 100 FOV (148 μm^2) images per lane or on a high-throughput single-molecule imaging system at Epivision obtaining 100 FOVs (200 μm^2) images per lane. The focus was maintained throughout the run using IR-assisted autofocus system.

Single-molecule image analysis. Image analysis was performed with the open-source software CellProfiler (v.3.1.9; http://www.cellprofiler.org/) as previously described[28]. In brief, time-lapse images of antibody-binding events and TetraSpeck beads were aligned, stacked and summed to one image. Antibody spots can be differentiated from TetraSpeck bead spots based on spot size and intensity. Summed antibody images were aligned with cell-free nucleosome images to count colocalization events. For biomarkers analysis, antibody binding events were quantified. To minimize the surface batch effects, signals were normalized to the surface activity (average antibody signal across all lanes of the surface).

EPINUC sample classification and feature scoring. PPS analysis on the data was conducted using a previously published algorithm (https://github.com/8080labs/ppscore). In brief, by calculating a cross-validated decision tree (repeated (10,000) 4-fold cross-validation) for the target variable (for example, diagnosis) using only one of the markers, it is possible to determine which of the markers in the datasets contributes most to the target variable.

XGBoost classifier was applied to the data, using all features measured and their combinations. Feature contribution was assessed by SHAP.

For ROC curves, for each classification task, the most important features were selected using the Boruta algorithm (with a random forest classifier)[121]. An additional feature engineering step was used to generate combinatorial combinations (multiplication, division and SQRT) of the selected features, and another round of feature selection was performed using Boruta. The final features selected were used either in a random forest classifier or in a soft voting classifier consisting of RF, XGBoost and LDA classifiers. The final classifier was used to generate a ROC curve with the associated confidence intervals using 10-fold cross-validation and 1,000 bootstrap repeats.

To enable comparison across independently processed EPINUC cohorts, batch correction was applied to account for differences in antibody labelling degree. In brief, the reference dataset was subsampled to balance group sizes, and feature-wise means and standard deviations were calculated. The independently processed cohort was first standardized and subsequently rescaled to match the mean and variance of the reference dataset. This approach allowed alignment of feature distributions across cohorts while preserving relative differences between samples.

Histone PTMs quantification by mass spectrometry

Histones were enriched from FFPE tissues as previously described[122]. Approximately 4 μg of histone octamer was mixed with an equal amount of heavy-isotope labelled histones, which were used as an internal standard[123], and separated on a 17% SDS–PAGE gel. Histone bands were excised, chemically acylated with deuterated acetic anhydride and in-gel digested with trypsin[124]. The samples were separated by

reversed-phase chromatography, acquired on a Q Exactive HF instrument (Thermo Fisher Scientific), and analysed using MaxQuant and QualBrowser, as previously described[124]. For each histone modified peptide, a percentage relative abundance (%RA) value for the sample (light channel) or the internal standard (heavy channel) was estimated. Light to heavy ratios (log_2 transformed) of %RA were then calculated and are reported in Supplementary Table 8.

Reporting summary

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

Data availability

Raw CyTOF and EPINUC measurements are available at Zenodo (https://doi.org/10.5281/zenodo.15336152)[125]. Owing to protection of personal data requirements, the patient-derived sequencing data generated in this study have been deposited into the European Genome-phenome Archive, under controlled access, with accession ID EGAS50000002070. Access is subject to approval by the Weizmann Breast Cancer Epigenetics Data Access Committee, consistent with participant consent and ethical approvals. The MARS-seq and H3K4me1 CUT&Tag data of sorted cells have been deposited into the Gene Expression Omnibus under accession GSE334437. The MS histone PTM data have been deposited into the ProteomeXchange Consortium via the PRIDE[126] partner repository with the dataset identifier PXD057150. Publicly available data and resources used in this study include the TCGA BRCA cohort (dbGaP Study Accession phs000178, Firehose stddata_2016_01_28, https://doi.org/10.7908/C11G0KM9), a breast cancer single-cell multiome dataset[13] (GSE243526), JASPAR transcription-factor binding profiles[119] and scRNA-seq data of human normal mammary epithelial cells[50] (GSE180878). Source data are provided with this paper.

Code availability

The code used for the analyses of WGBS, snCUT&Tag and EPINUC data is available at Zenodo (https://doi.org/10.5281/zenodo.15336152)[125]. The CyTOF analysis code is available at Zenodo (https://doi.org/10.5281/zenodo.21982487)[127].

89. Ben-David, U. et al. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* **560**, 325–330 https://doi.org/10.1038/s41586-018-0409-3 (2018).
90. Bankhead, P. et al. QuPath: open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 https://doi.org/10.1038/s41598-017-17204-5 (2017).
91. Pachitariu, M., Rariden, M. & Stringer, C. Cellpose-SAM: superhuman generalization for cellular segmentation. Preprint at *bioRxiv* https://doi.org/10.1101/2025.04.28.651001 (2025).
92. Li, M. M. et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J. Mol. Diagn.* **19**, 4–23 https://doi.org/10.1016/j.jmoldx.2016.10.002 (2017).
93. Bagwell, C. B. et al. Automated data cleanup for mass cytometry. *Cytometry A* **97**, 184–198 https://doi.org/10.1002/cyto.a.23926 (2020).
94. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2020).
95. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. Second International Conference on Knowledge Discovery and Data Mining* 226–231 (AAAI Press, 1996).
96. Friedman, J. & Dieng, A. B. 2023. The Vendi Score: a diversity evaluation metric for machine learning. Preprint at https://doi.org/10.48550/arXiv.2210.02410 (2023).
97. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 https://doi.org/10.1038/s41592-019-0638-x (2020).
98. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 https://doi.org/10.1089/omi.2011.0118 (2012).
99. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 https://doi.org/10.14806/ej.17.1.200 (2011).
100. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232 https://doi.org/10.1186/1471-2105-10-232 (2009).
101. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 https://doi.org/10.1101/gr.107524.110 (2010).
102. Sun, D. et al. MOABS: model based analysis of bisulfite sequencing data. *Genome Biol.* **15**, R38 https://doi.org/10.1186/gb-2014-15-2-r38 (2014).
103. Hetzel, S., Giesselmann, P., Reinert, K., Meissner, A. & Kretzmer, H. RLM: fast and simplified extraction of read-level methylation metrics from bisulfite sequencing data. *Bioinformatics* **37**, 3934–3935 https://doi.org/10.1093/bioinformatics/btab663 (2021).
104. Jaitin, D. A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
105. Kohen, R. et al. UTAP: user-friendly transcriptome analysis pipeline. *BMC Bioinformatics* **20**, 154 https://doi.org/10.1186/s12859-019-2728-2 (2019).
106. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* **27**, 1571–1572 https://doi.org/10.1093/bioinformatics/btr167 (2011).
107. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 https://doi.org/10.1093/bioinformatics/btv145 (2015).
108. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 https://doi.org/10.1093/nar/gkh103 (2004).
109. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 https://doi.org/10.1093/bioinformatics/btq033 (2010).
110. Onuchic, V. et al. Epigenomic deconvolution of breast tumors reveals metabolic coupling between constituent cell types. *Cell Rep.* **17**, 2075–2086 https://doi.org/10.1016/j.celrep.2016.10.057 (2016).
111. Korotkevich, G. et al. Fast gene set enrichment analysis. Preprint at *bioRxiv* https://doi.org/10.1101/060012 (2021).
112. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 https://doi.org/10.1073/pnas.0506580102 (2005).
113. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat. Biotechnol.* **39**, 825–835 https://doi.org/10.1038/s41587-021-00869-9 (2021).
114. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 https://doi.org/10.1038/s41592-021-01282-5 (2021).
115. Hao, Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 https://doi.org/10.1038/s41587-023-01767-y (2023).
116. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 https://doi.org/10.1093/bioinformatics/btr260 (2011).
117. Andreatta, M. & Carmona, S. J. UCell: robust and scalable single-cell gene signature scoring. *Comput. Struct. Biotechnol. J.* **19**, 3796–3798 https://doi.org/10.1016/j.csbj.2021.06.043 (2021).
118. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 https://doi.org/10.1186/gb-2008-9-9-r137 (2008).
119. Castro-Mondragon, J. A. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 https://doi.org/10.1093/nar/gkab1113 (2021).
120. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 https://doi.org/10.1038/s41587-022-01284-4 (2022).
121. Kursa, M. B. & Rudnicki, W. R. Feature selection with the Boruta package. *J. Stat. Softw.* **36**, 1–13 https://doi.org/10.18637/jss.v036.i11 (2010).
122. Noberini, R., Restellini, C., Savoia, E. O. & Bonaldi, T. Enrichment of histones from patient samples for mass spectrometry-based analysis of post-translational modifications. *Methods* **184**, 19–28 https://doi.org/10.1016/j.ymeth.2019.10.001 (2020).
123. Noberini, R. & Bonaldi, T. A super-SILAC strategy for the accurate and multiplexed profiling of histone posttranslational modifications. *Methods Enzymol.* **586**, 311–332 https://doi.org/10.1016/bs.mie.2016.09.036 (2017).
124. Noberini, R. et al. PAT-H-MS coupled with laser microdissection to study histone post-translational modifications in selected cell populations from pathology samples. *Clin. Epigenetics* **9**, 69 https://doi.org/10.1186/s13148-017-0369-8 (2017).
125. Ron, G., Furth, N., & Shema, E. Epigenetic heterogeneity in breast tumors with implications for disease diagnosis and monitoring. *Zenodo* https://doi.org/10.5281/zenodo.15336152 (2025).
126. Perez-Riverol, Y. et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **50**, D543–D552 https://doi.org/10.1093/nar/gkab1038 (2022).
127. Ron, G. CyTOF analysis code for: epigenetic heterogeneity in breast tumors with implications for disease diagnosis and monitoring (version v1.0.0) [computer software]. *Zenodo* https://doi.org/10.5281/zenodo.21982487 (2026).

Acknowledgements We thank U. Ben-David for sharing MCF7 single-cell-derived clones; I. Ben-Porath for sharing HCC70, HCC1937 and SUM149 cell lines; and V. Holiar, S. Lebon and M. Biton for their assistance with cell sorting.

Author contributions N.F., G.R. and E.S. designed the study. N.F. and C.R. processed tumour samples and performed CyTOF experiments. Y.A. established the CyTOF analysis framework. C.R. conducted all cell culture experiments. M.A., M.-Y.P., H.Y., N.E., E.A., A.S., J.Z. and D.J. contributed to EPINUC analyses. S.K., O.A. and T.G. contributed to MS analysis of sorted tumour cells. T.M.S. assisted with CyTOF experiments. R.N. and T.B. contributed to MS analysis of histone PTMs. A.D., G.J. and C.V. contributed to snCUT&Tag analyses. B.D., R.T., G.S. and A.P.F. contributed to WGBS analysis. O.G. assisted with image analyses. M.H. and S.D.M. performed single-cell multiome analyses. M.H., A.K., S.D.M. and A.M. contributed to EM-seq analysis of sorted cells. M.Z. and R.O. provided PDX models. I.M. conducted the TruSight Oncology 500 assay. B.B.-Z., Y.M., O.F.-E., H.K.-S., S.G., O.G.-M., N.R., F.A.-R., S.I.-S., A.Y., A.A., S.M., M.K., M.B., M.D. and E.N.G.-Y. provided tumour and blood samples. N.F., Y.A., C.R., E.S. and G.R. analysed datasets. N.F., G.R. and E.S. wrote the manuscript.

Funding C.R. was supported by a postdoctoral fellowship from the Azrieli Foundation. This research was supported by grants from the European Research Council (ERC101115455 to E.S. and ERC101044574 to T.G.), The Israeli Science Foundation (1349/24 to E.S.), the German-Israeli Cooperation in Cancer Research (to E.S. and G.R.), Israel Cancer Research Fund-Diane and a Guilford Glazer Foundation Acceleration Grant (to E.S. and G.R.), ERA-NET Sustained collaboration of national and regional programmes in cancer research (to E.S.), TRANSCAN-3 Joint Transnational Call (EPIPROPER, to E.S., T.B. and C.V.), an Emerson Collective, Minerva Research grant, the Swiss Society Institute for Cancer Prevention Research, Magnus Konow in Honor of his mother Olga Konow Rappaport and the Fundación Alberto Palatchi (to E.S.).

Competing interests Yeda Research and Development has filed a US Provisional patent application US-20240352538-A1: 'Methods and Kits for Analyzing Nucleosomes and Plasma Proteins', which are related to Figs. 4 and 5 of this paper, and describe the isolation and imaging of cell-free nucleosomes from plasma (the EPINUC technology). E.S., N.E., A.S., K.A. and D.J. are named inventors. Epivision Genetics holds exclusive licence to the patent application related to this work. The remaining authors declare no competing interests.

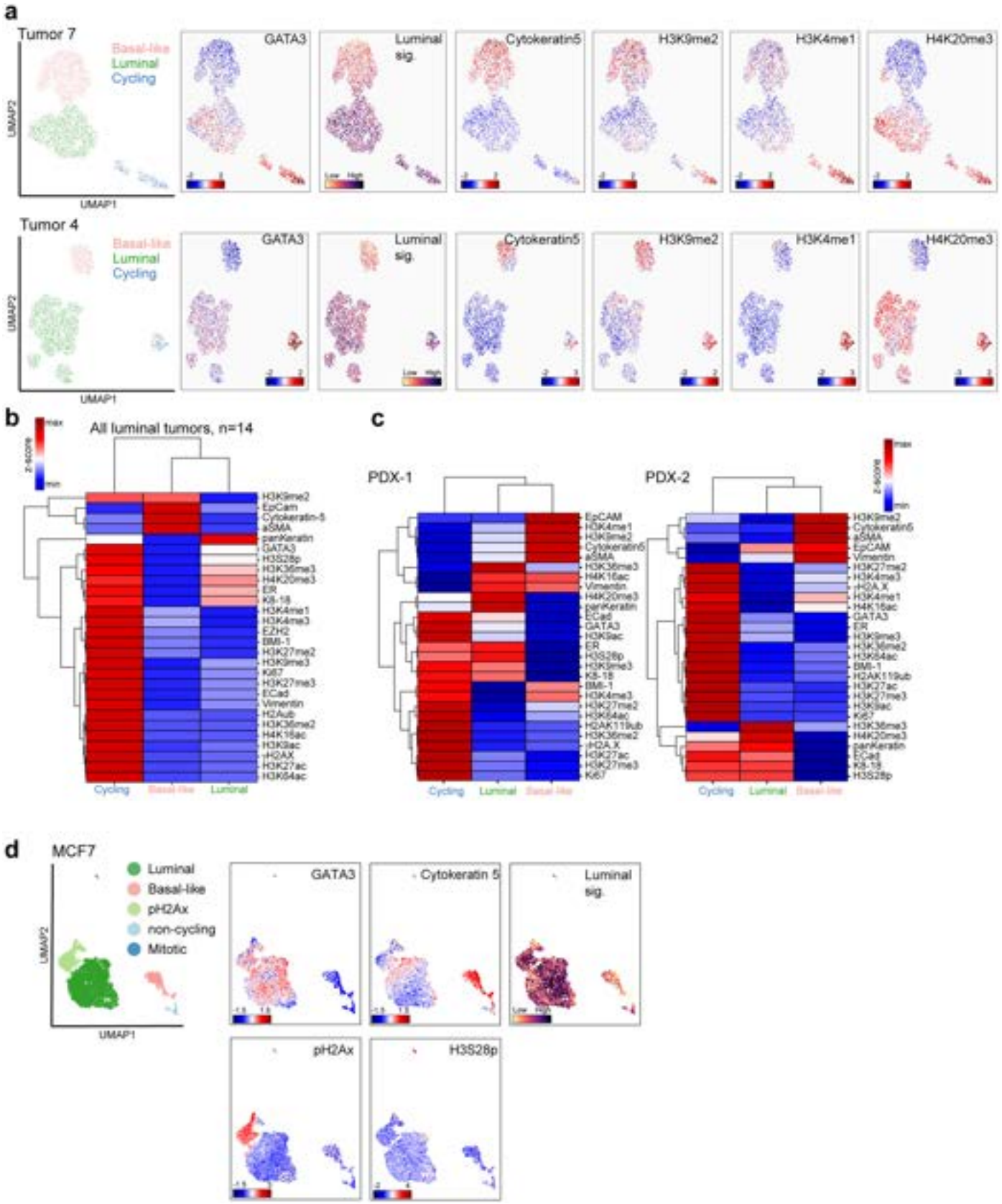
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-026-11230-y.

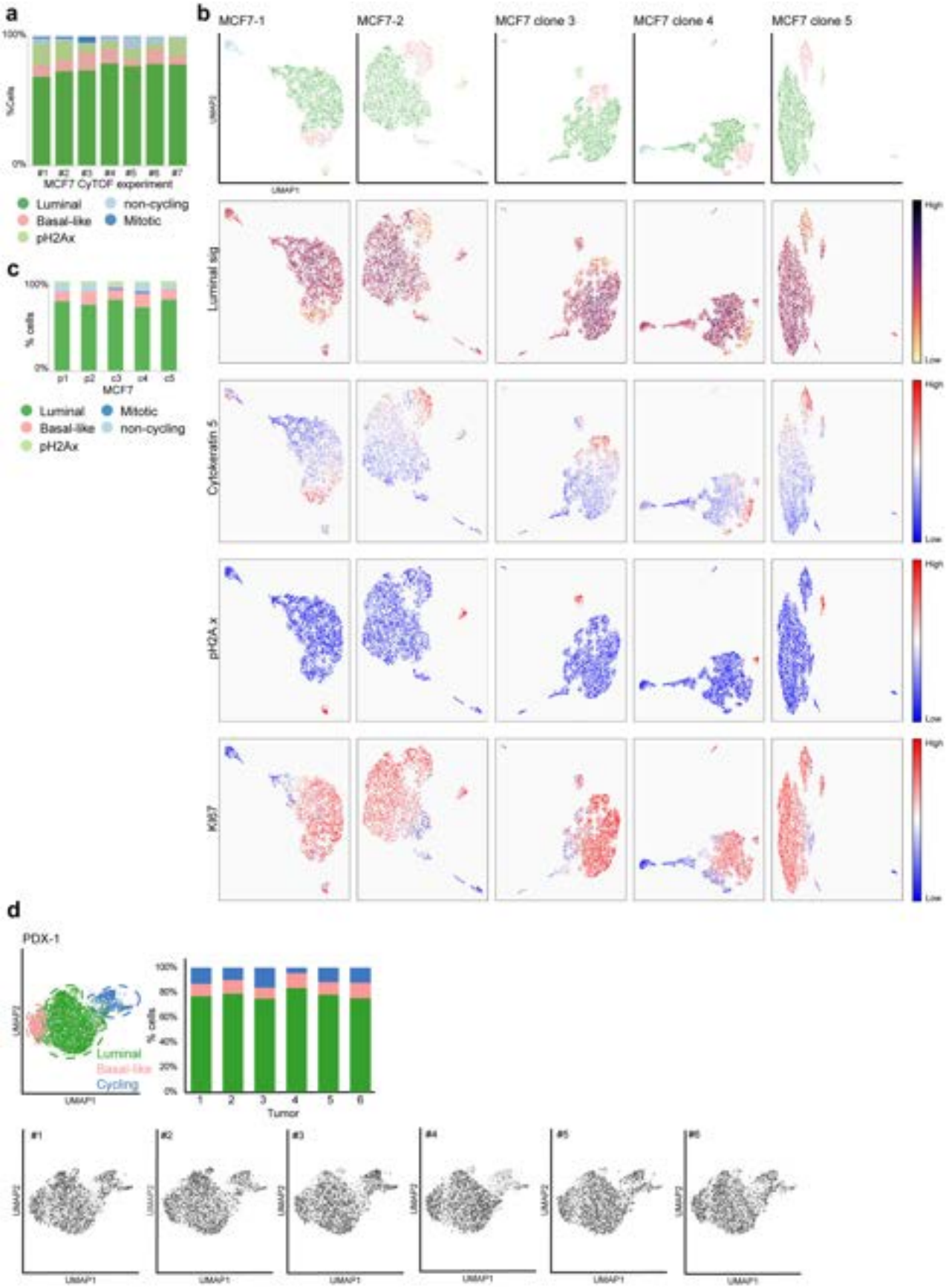
Correspondence and requests for materials should be addressed to Guy Ron or Efrat Shema.

Peer review information *Nature* thanks Michail Ignatiadis, who co-reviewed with Jana Jeschke; Lluis Morey, who co-reviewed with Liliana Garcia Martinez; Heather Parsons; and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at http://www.nature.com/reprints.

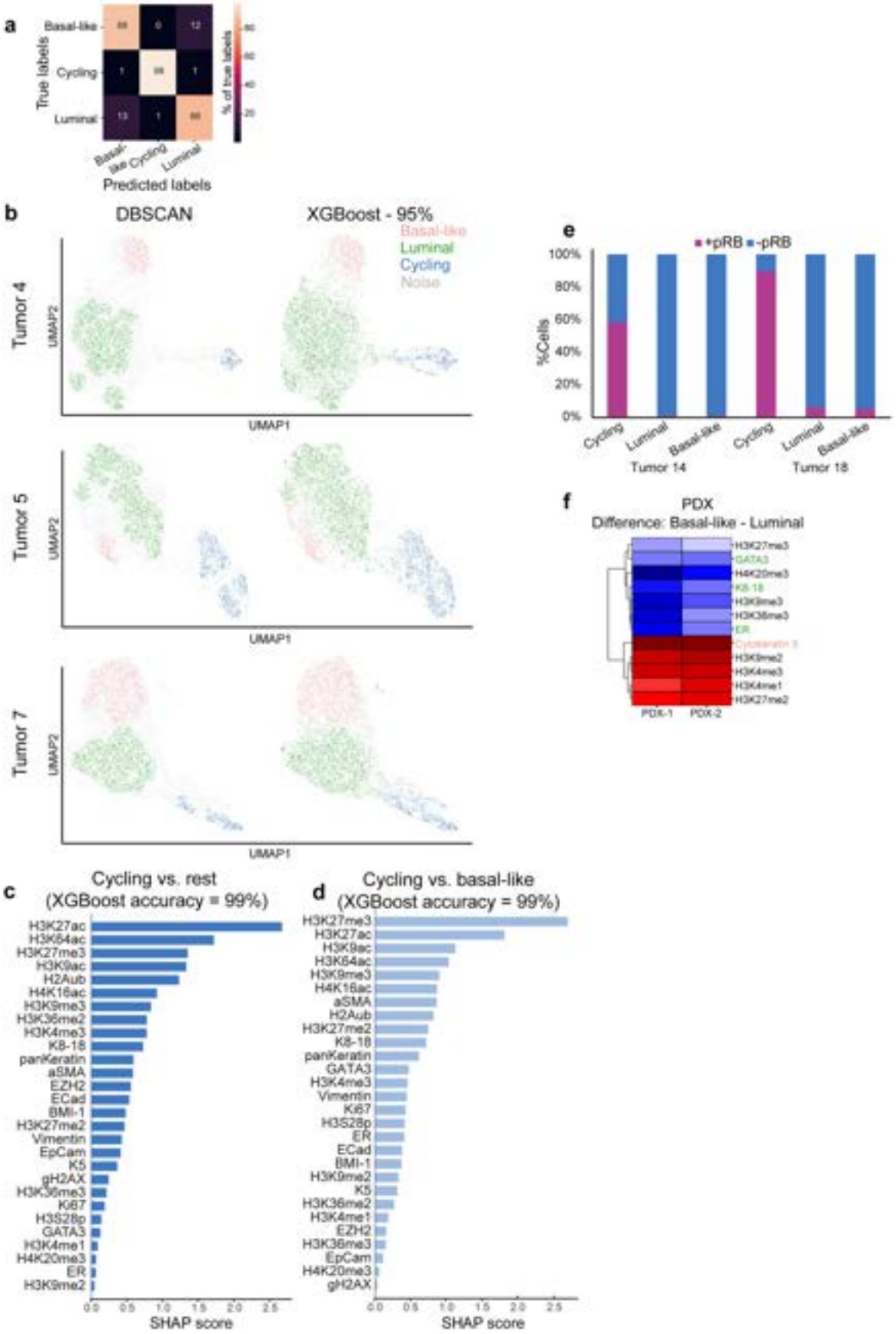


Extended Data Fig. 1 | Tumor cell classification into Cycling, basal-like and luminal subpopulations. (a) Tumors 4 and 7 analyzed as in Fig. 1h. Cells are colored by classification or the indicated feature levels. (b) Heatmap of mean feature levels in each subpopulation across all HR+ tumors analyzed by CyTOF (n = 14, including 3 anchor repeats, see Table S3). Rows present standardized values. (c) Two luminal PDX models analyzed by CyTOF after removal of mouse cells (human CD298-positive; mouse MHC class 1-negative), and classification into three subpopulations (Fig. 1j). PDX-1 comprised of six tumors, with equal cell sampling per tumor; PDX-2 comprised of one tumor. Heatmaps show the mean feature levels within each subpopulation for each model. (d) UMAP of MCF7 CyTOF data using all measured markers. Points are colored by cluster (left) or the indicated feature or signature (right).



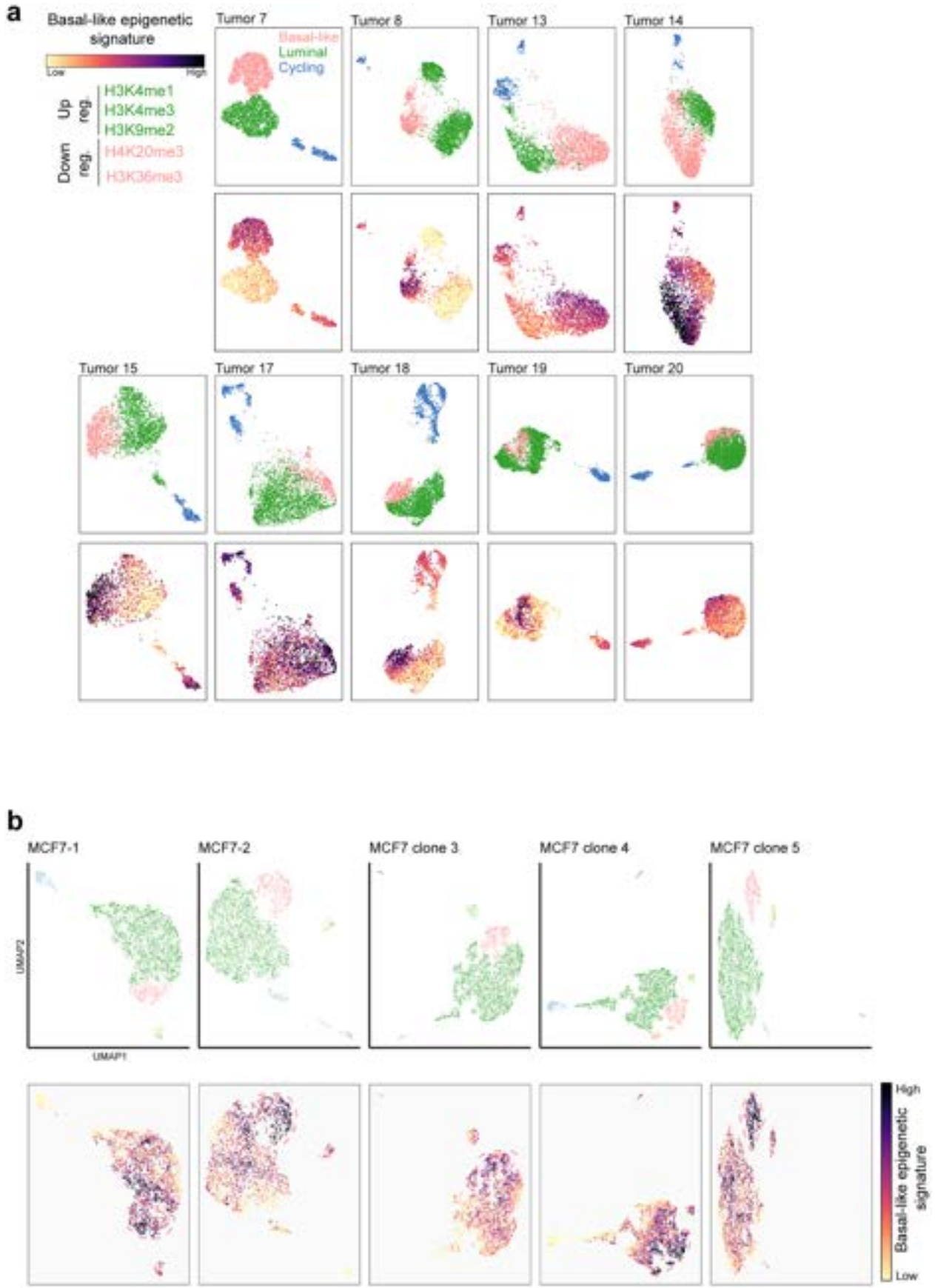
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | CyTOF analysis of MCF7 cells and luminal PDX models confirms cell-intrinsic epigenetic states. (a) Distribution of DBSACN-defined cell subpopulations across independent MCF7 CyTOF experiments of MCF7 cells, showing consistent proportions of the different cell states. (b) UMAP representation and DBSCAN clustering of two different MCF7 parental cells or single-cell derived clones analyzed by CyTOF, using only epigenetic features. Points are colored by the classification of different subpopulations or the expression levels of the indicated feature or signature (PermCell). (c) The distribution of cells among the populations defined by DBSCAN clustering in the MCF7 parental (p) or single-cell derived clones (c). The single-cell derived clones recapitulate the entire spectrum of heterogeneity observed in the parental population. (d) UMAP representation of CyTOF analysis of six independently grown tumors from the PDX-1 model (n = 6 tumors, each grown in a different mouse), colored by subpopulations. Cells from each individual tumor are mapped onto the joint UMAP, and the distribution of cells among the three subpopulations for each tumor is shown.

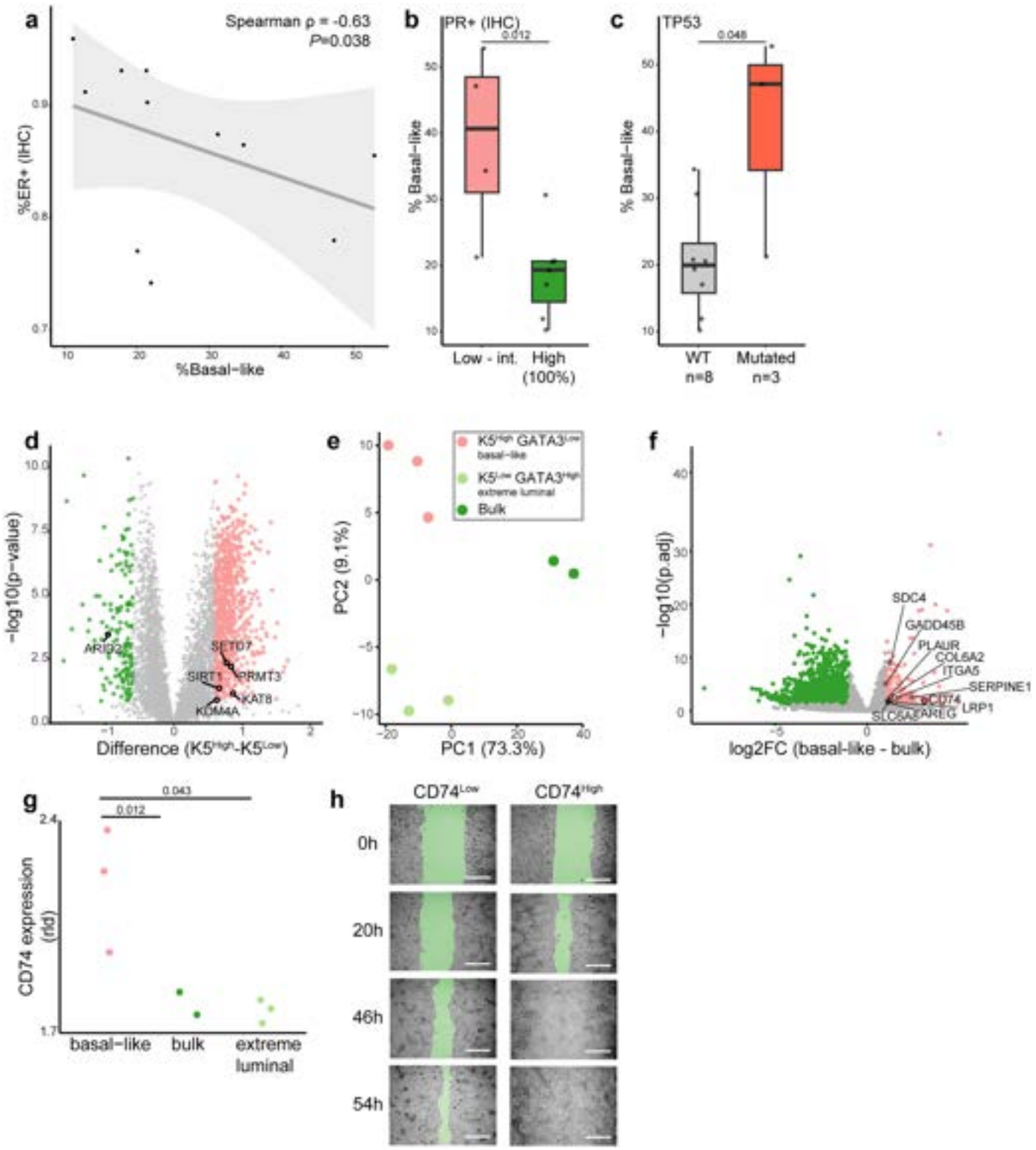


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Machine-learning-based classification of epigenetic distinct sub-populations within luminal breast tumors. (a) Confusion matrix of XGBoost model predictions for all 3 subpopulations across HR+ tumors (n = 14) using only epigenetic features, excluding cell identity markers. Rows were normalized to show predictions as percentages of cells within each true class. (b) Representative examples of XGBoost-based noise prediction procedure. UMAPs of individual tumors are shown before (left) and after (right) prediction. Points are colored by classification; non-clustered cells before prediction, or cells in which prediction confidence <95% are marked in gray. (c-d) XGBoost models distinguishing Cycling cells from all other tumor cells (**c**) or specifically the basal-like cells (**d**) within the HR+ CyTOF cohort. SHAP scores show feature contributions to classification, ranked in descending order. (e) Percentage of proliferating cells (assessed by the levels of pRb, as in Fig. S3) in each epigenetic sub-populations (Cycling, luminal and basal-like). The basal-like and luminal sub-populations do not differ due to differences in the proliferative state. (f) The differences in the average levels of the indicated features between the basal-like and luminal sub-populations in the luminal PDX models. Basal-like cells show a recurrent pattern of both up and downregulated histone PTMs, highly similar to the pattern observed in fresh human tumors.

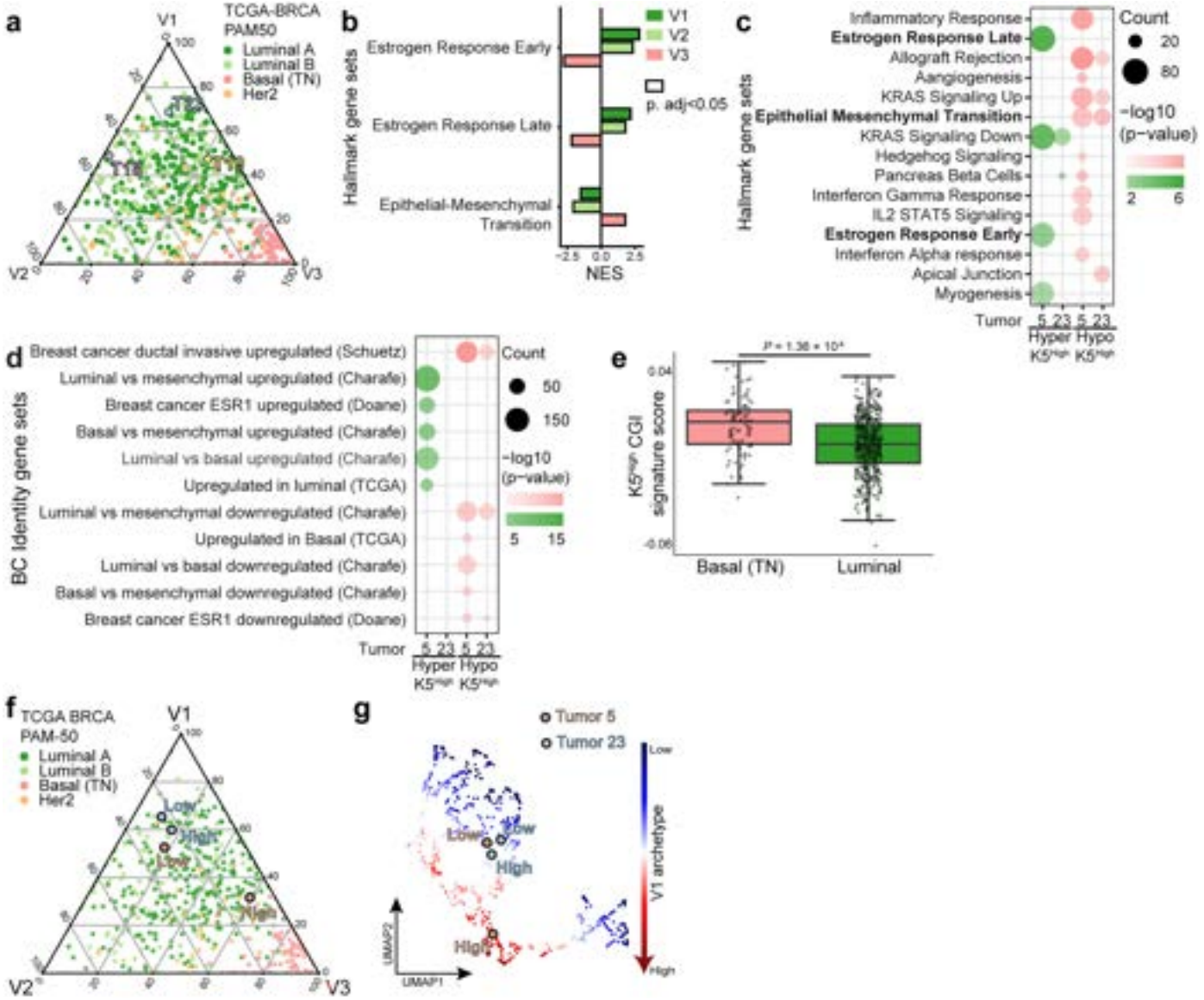


Extended Data Fig. 4 | Epigenetic basal-like PermCell score for the identification of basal-like cells. (a) UMAP of CyTOF data from tumors (as in Fig. S4), colored by the assigned clusters (top) or the basal-like epigenetic signature (basal PermCell score, bottom), which is derived solely from histone modifications. (b) UMAP of two MCF7 parental cultures and three single-cell-derived clones analyzed by CyTOF using only epigenetic features, colored by assigned clusters (top) or the basal-like epigenetic signature (bottom).

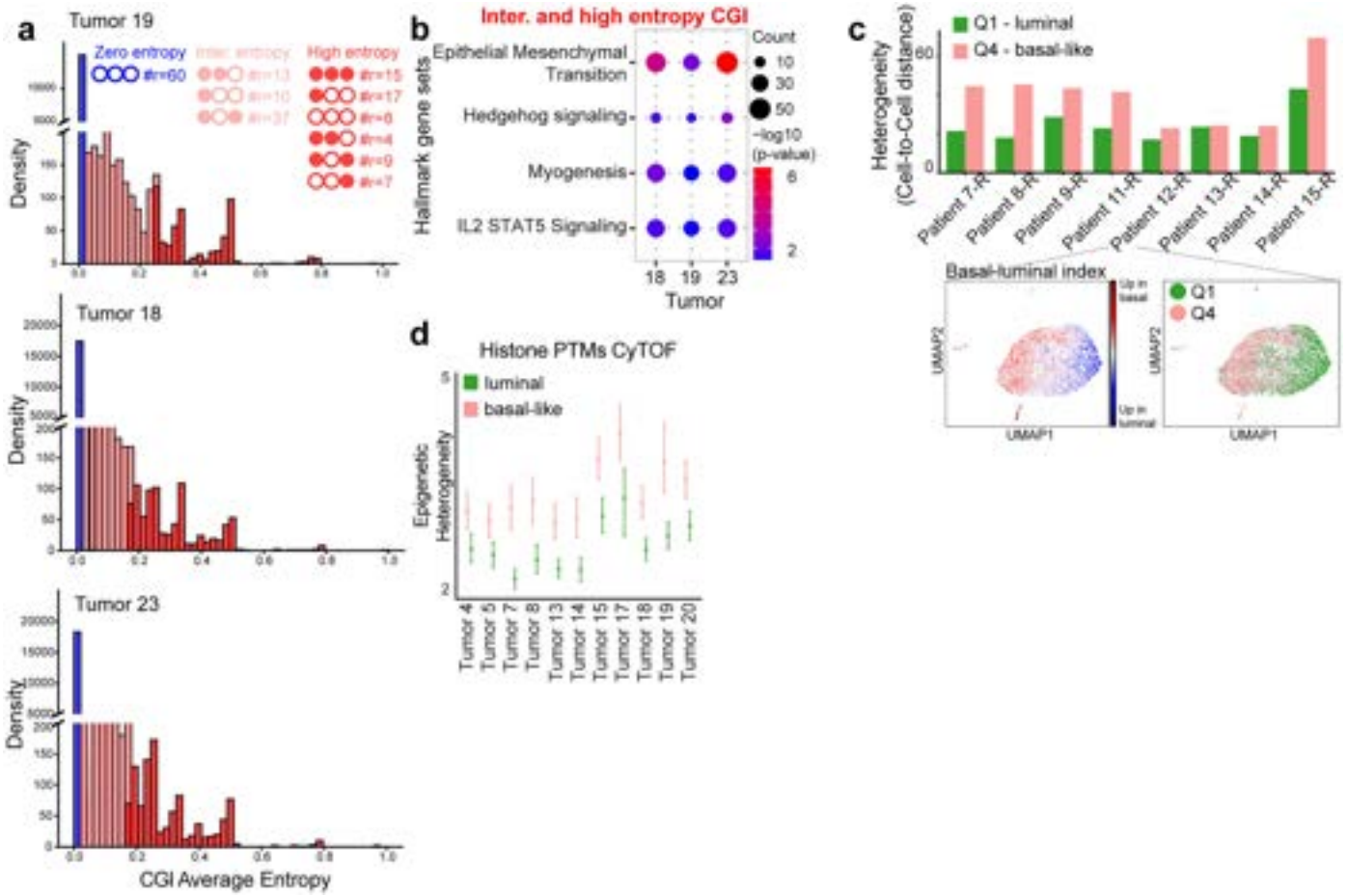


Extended Data Fig. 5 | See next page for caption.

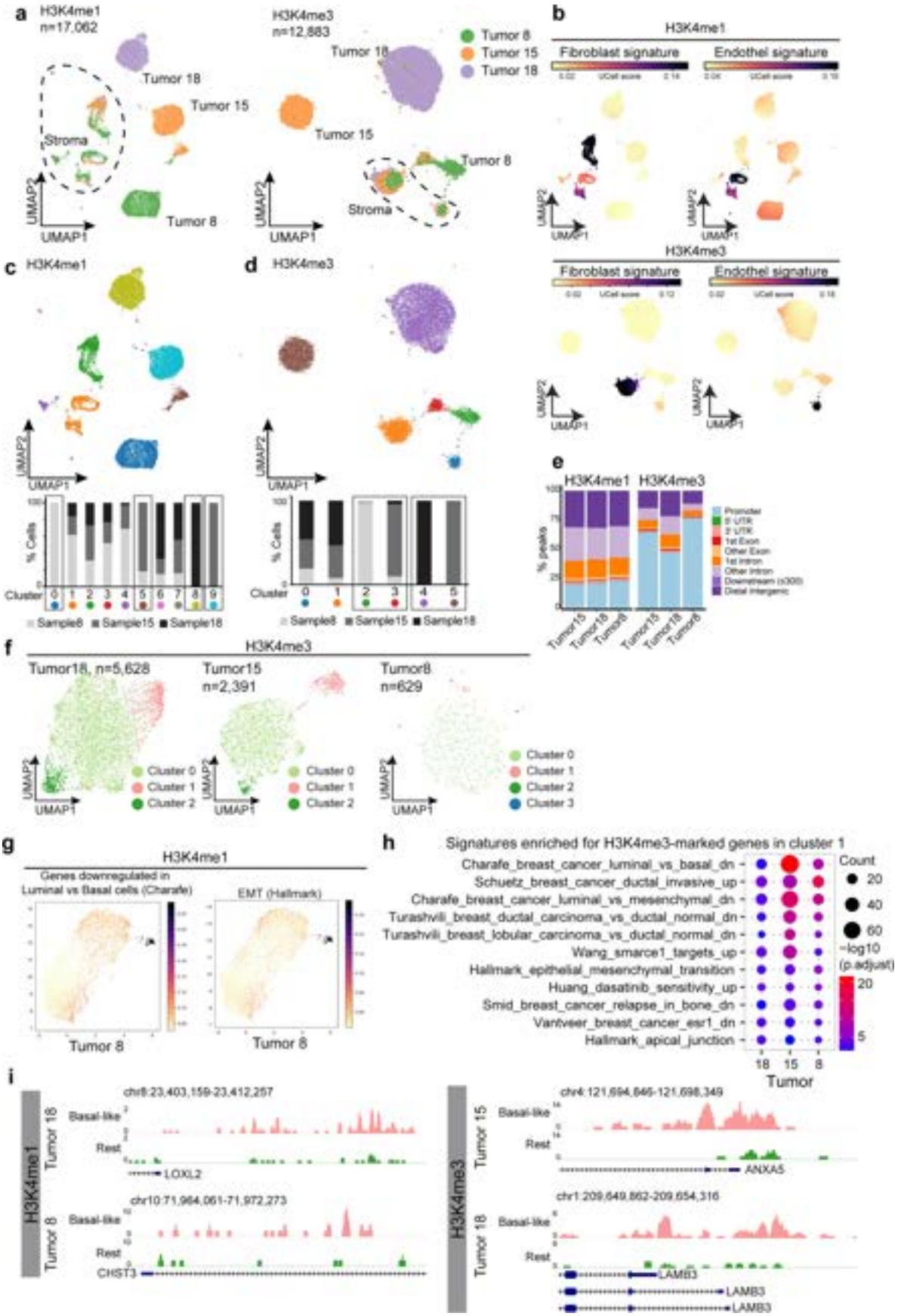
Extended Data Fig. 5 | Functional characterization of the basal-like population. (a) Scatter plot comparing basal-like cell fraction with estrogen receptor (ER) positivity by IHC across tumors (n = 11). Spearman correlation coefficient and two-sided p-value are indicated. (b,c) Basal-like cell fraction in HR+ tumors stratified by progesterone receptor (PR) IHC (b; 100% PR-positive, n = 7; lower PR positivity, n = 4) or TP53 status (c; wild-type, n = 8; mutant, n = 3). Two-sided Wilcoxon tests. Boxes show median and 25th–75th percentiles; whiskers extend within 1.5xIQR. (d) Differential protein abundance between $K5^{High}$ and $K5^{Low}$ populations in tumor 19. Axes show protein intensity difference ($K5^{High}$-$K5^{Low}$) and -log10(p-value), respectively. Two-sided Student's t-tests with FDR correction; Green and pink dots depict proteins significantly deregulated (Fold change>1.5, q-value < 0.05). Labeled are chromatin regulators differentially expressed between the two populations. (e) Principal component analysis (PCA) of RNA-seq profiles from sorted MCF7 subpopulations. Basal-like cells ($K5^{High}$ $GATA3^{Low}$, n = 3), extreme luminal cells ($K5^{Low}$ $GATA3^{High}$, n = 3), and bulk MCF7 cells (n = 2). Points represent independently sorted population. Axes indicate variance explained. (f) Differential gene expression between basal-like ($K5^{High}$ $GATA3^{Low}$, n = 3) and bulk MCF7 cells (n = 2). Significantly upregulated (pink) and downregulated (green) genes are indicated (DESeq2, Benjamini–Hochberg-adjusted p-value < 0.05, log2FC > 1). EMT-associated genes (hallmark epithelial-mesenchymal transition gene set) are enriched among upregulated transcripts in the basal-like population and highlighted with a black outline. CD74, a membrane-bound protein upregulated in the basal-like cells, is marked in dark pink. (g) CD74 expression across basal-like (n = 3), bulk (n = 2), and extreme luminal (n = 3) MCF7 populations. Points represent independently sorted population; expression is shown as rld values. Statistical significance was derived from DESeq2 differential expression analysis with Benjamini–Hochberg correction for multiple testing. (h) Representative wound-healing assay images of sorted $CD74^{Low}$ and $CD74^{High}$ MCF7 cells at the indicated times. Three independently recovered sorted populations from the same sort were analyzed. The gap area is highlighted in green. Scale bar = 300 μm.



Extended Data Fig. 6 | Functional analysis of deconvoluted WBGS data of luminal breast tumors. (a) TCGA-BRCA DNA-methylation dataset (n = 782)[51] was deconvoluted into three archetypes. Dots show tumor archetype composition and colored by PAM50 subtype. The archetype V3 is associated with basal-like triple-negative tumors. The composition of archetypes within the three tumors analyzed in this study (T18, T19, and T23, highlighted) clusters with luminal breast tumors from the TCGA cohort. (b) Genes were ranked by correlation between methylation archetypes and matched TCGA-BRCA expression, followed by GSEA[111] against MSigDB hallmark gene sets (Table S4b). V1 and V2 archetypes are associated with luminal transcriptional programs while V3 archetype is enriched for the EMT program. Normalized Enrichment Scores (NES) of representative terms are shown. Black outline denotes BH-adjusted p-value < 0.05. See Fig. S5a for all terms. (c-d) $K5^{High}$ and $K5^{Low}$ sorted cells from tumors 5 and 23 were subjected to enzyme-mediated methylation sequencing (EM-seq, n = 4 replicates, 200 cells/replicate), and differentially methylated gene promoters were identified. For each sample, functional enrichment analysis was performed separately for genes with hypermethylated or hypomethylated promoters. Shown are the top 10 enriched terms from the MSigDB Hallmark gene set collection (**c**) or a curated list of breast cancer cell identity signatures (**d**) with nominal p-value < 0.05 (one-sided hypergeometric test). Circle size represents the number of genes from the input list overlapping each gene set, and transparency reflects -log10(p-value). (e) A CpG-island (CGI) methylation signature distinguishing $K5^{High}$ and $K5^{Low}$ states, derived using CGIs showing concordant differential methylation between the states in both tumor 5 and 23. Shown is the projection of this signature into DNA methylation profiles from the TCGA-BRCA dataset. Positive scores indicate greater similarity to the EM-seq-derived $K5^{High}$ methylation program, whereas negative scores indicate similarity to the $K5^{Low}$ state. Boxplots show the distribution of signature scores across basal-like (n = 96) or luminal (n = 418) tumors with points representing individual tumors. Boxes show median and 25th–75th percentiles; whiskers extend to the largest/smallest value within 1.5×IQR. The $K5^{High}$ methylation program is enriched in basal-like breast tumors. Two-sided Mann–Whitney U test. (f) TCGA-BRCA DNA-methylation dataset deconvoluted into three archetypes as in **a**, with $K5^{high}$ and $K5^{low}$-sorted populations from tumors 5 and 23 projected onto the archetype space. $K5^{high}$ populations are shifted toward the V3 archetype, consistent with enrichment for basal/EMT-associated methylation programs. (g) Representation of tumors from the TCGA-BRCA cohort based on the four-archetype decomposition described in Fig. 3a, colored by V1 contribution. $K5^{high}$ and $K5^{low}$-sorted populations from tumors 5 and 23 are projected onto the UMAP embedding. $K5^{high}$ populations are localized to regions characterized by higher V1 contribution, consistent with the EMT-associated methylation programs.

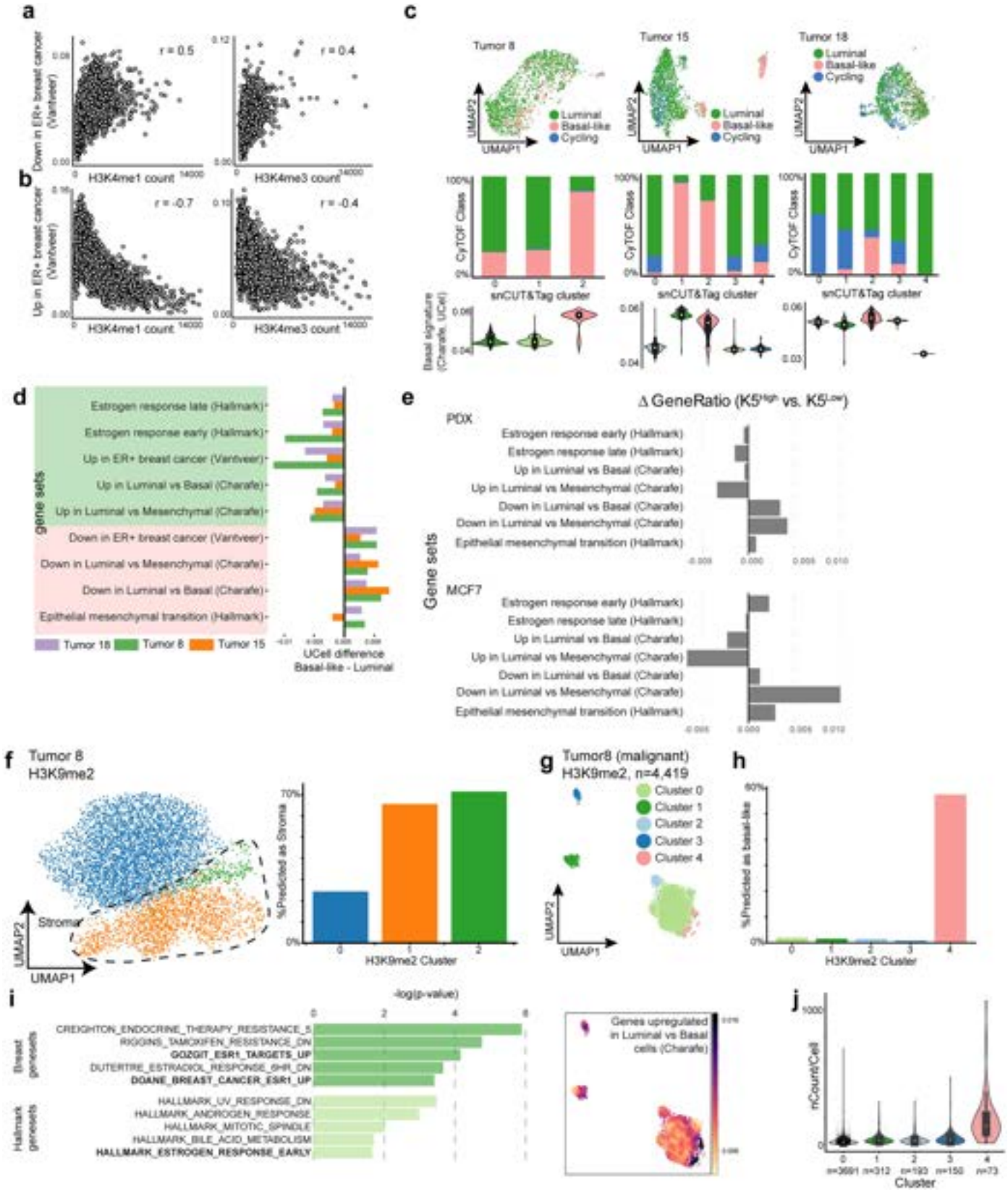


Extended Data Fig. 7 | Basal-like cells are associated with increased epigenetic and transcriptional heterogeneity. (a) WGBS reads were clustered by DNA methylation patterns over 100 bp bins[55], and each bin was scored for entropy, reflecting DNA methylation pattern heterogeneity; bins containing one pattern have zero entropy. Illustration, rows represent methylation patterns, filled/open circles indicate methylated/unmethylated CpGs and #r denotes supporting reads. Histograms show mean entropy scores of CGIs in tumor 19 (n = 17,434), tumor 18 (n = 20,845) and tumor 23 (n = 22,440). Most CGIs are unmethylated and have zero entropy scores. The intermediate and high-entropy bins are marked in pink and red. (b) Intermediate- and high-entropy CGIs (entropy>0) were linked to genes and analyzed for enrichment against MSigDB Hallmark genesets[85]. Pathways significantly enriched in all three tumors are shown (one-sided hypergeometric test, nominal p-value < 0.05). Dot size indicates overlapping genes; color indicates -log10(p-value). Variable CGIs are significantly enriched for genes related to EMT. (c) The Basal-luminal index was used to score the transcriptional identity of tumor cells[13], and cells from each tumor were divided into quartiles; a low index score corresponds to expression of luminal identity genes (Q1) and a high index score corresponds to expression of basal-like identity genes (Q4). Cell-to-cell distance was calculated within each quartile to assess transcriptional heterogeneity; medians are shown for each sample. Bottom, representative UMAP and quartile definition for tumor 12-R. (d) Vendi score[96] was used to calculate epigenetic heterogeneity within luminal or basal-like subpopulation (as defined in Figs. 1, 2, across n = 11 luminal tumors) using single-cell CyTOF histone PTMs measurements. Points show the mean across 5,000 rarefactions; error bars indicate uncertainty intervals (2.5th–97.5th percentiles) of the rarefied score distribution. Histone PTM levels are more variable in basal-like cells.



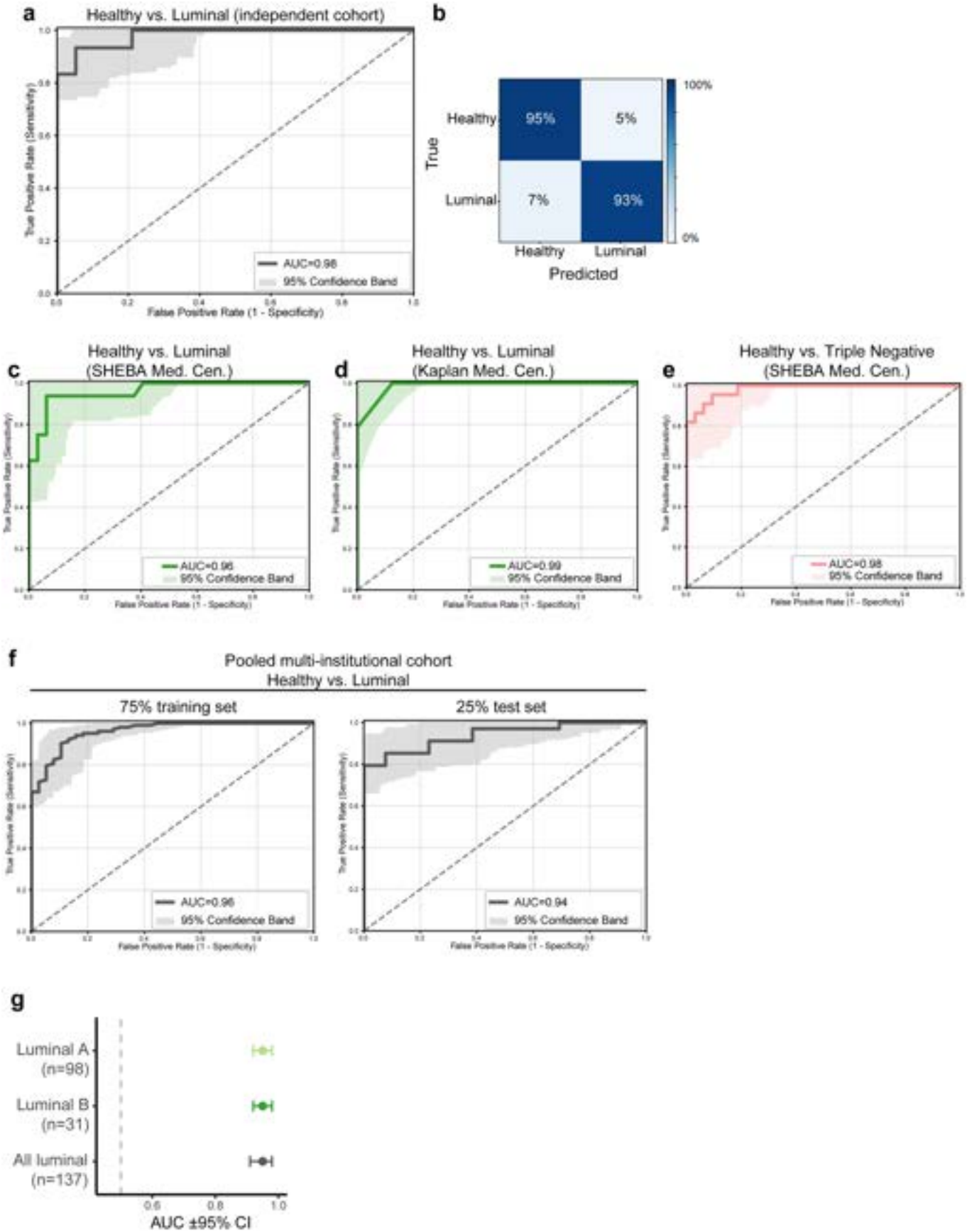
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | H3K4me1/3 snCUT&Tag of three luminal tumors. (a) UMAPs of H3K4me1 (left, n = 17,062 cells) and H3K4me3 (right, n = 12,883 cells) snCUT&Tag 50 kb bins matrices from three tumor samples (PC1 representing read counts excluded; CD45-negative cells). Shared stroma clusters are outlined. (b) UMAPs as in **a**, colored according to H3K4me1 or H3K4me3 enrichment over fibroblast identity genes: FAP, COL3A1 and DCN (left) or endothelial identity genes: EMCN, PECAM1 and NOTCH3 (right). Higher UCell scores indicate stronger enrichment at genes associated with the corresponding stromal identity. (c,d) H3K4me1 (**c**) or H3K4me3 (**d**) UMAPS from all three tumors (50k bins matrix), colored by cluster. The portion of cells from each tumor that constitutes the clusters is shown below. Shared clusters were classified as stroma, as verified by gene activity; Boxed tumor clusters were analyzed further. For H3K4me3, limited number of cells passed quality control thresholds for tumor 8, thus data quality for this sample is lower. (e) Genomic annotation of H3K4me1 and H3K4me3 snCUT&Tag pseudo-bulk peaks, after stoma removal. Stacked bar plots show percentage of peaks assigned to genomic features. (f) For each tumor sample, removing the stroma compartment, a scTransform-normalized H3K4me3 gene activity was generated (excluding read counts information), represented by UMAP and clustered using Louvain clustering. Cell numbers are indicated. (g) Tumor 8 H3K4me1 UMAP (Fig. 3e), colored by UCell scores derived based on the genes contributing to the enrichment of the indicated terms. Cluster 2 shows high H3K4me1 signal on genes associated with basal-like cell identity. While the EMT signature was not significantly enriched for this cluster (FDR-adjusted p-value > 0.05), the corresponding UCell score is shown for comparison using EMT-associated genes contributing to the cluster. (h) Differentially H3K4me3-bound genes identified for each cluster shown in **f** (two-sided Wilcoxon signed-rank test, log2FC > 1, adjusted p-value < 0.01) were tested against MSigDB Hallmark and Breast-related genesets (Table S4). Terms enriched in cluster 1 in all three tumors are shown (one-sided hypergeometric test, FDR-adjusted P < 0.05). (i) Pseudu-bulk H3K4me1 and H3K4me3 genome-browser tracks from basal-like clusters (pink) or equal numbers of cells sampled from all other clusters (green) at representative genomic regions.



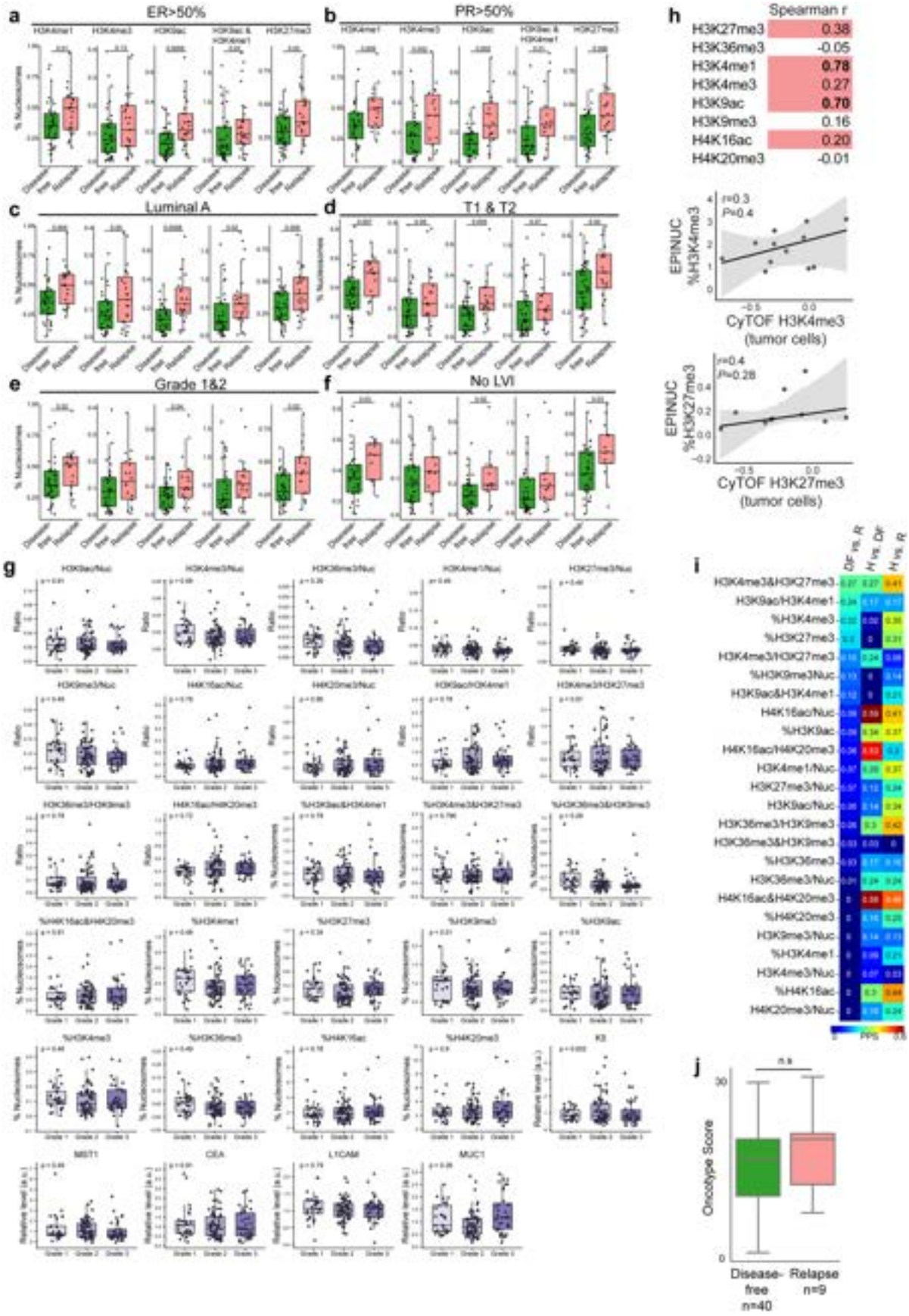
Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | CyTOF-snCUT&Tag integration and H3K9me2 genomic profiling. (a,b) H3K4me1 or H3K4me3 reads per cell versus negative ER-signaling score in tumor 15 (a) or positive ER-signaling score in tumor 18 (b). Points represent cells; r is Pearson's correlation coefficient. (c) H3K4me1 snCUT&Tag UMAPs from tumors 15 and 18 (Fig. 3e), colored by CyTOF-derived labels (top), with predicted-label composition per snCUT&Tag cluster (middle) and UCell scores for the 'Down in Luminal vs. Basal (Charafe)' signature (bottom). (d) Cells were classified as luminal or basal-like by CyTOF label transfer as in **c**. Differences in mean UCell scores for the indicated signatures between basal-like and luminal cells are shown across three tumors. EMT and basal-like-associated gene signatures show higher H3K4me1 deposition in the CyTOF-predicted basal-like group. (e) $K5^{High}$ and $K5^{Low}$ sorted cell populations from the luminal PDX model or MCF7 cells were subjected to H3K4me1 CUT&Tag. Peaks unique to each population were assigned to genes and compared to the indicated gene signatures. Shown is the difference in the GeneRatio (overlap count divided by the total number of peak-associated genes in that population) for each population, with positive values indicating increased overlap in $K5^{High}$ (basal-like) cells, whereas negative values indicate increased overlap in $K5^{Low}$ (luminal) cells. (f) Tumor 8 H3K9me2 snCUT&Tag UMAP from a 20kb-bin matrix, clustered by Leiden clustering (left). Stroma percentages per cluster, based on the H3K4me1-trained classifier, are shown on the right. (g) Following removal of the stromal compartment described in **f**, a normalized promoter-proximal H3K9me2 occupancy matrix was generated, and Leiden clustering was applied (n = 4,419 cells, see Methods). (h) Percentage of basal-like cells (cluster 2 in Fig. 3e) within each cluster in **g**, based on the H3K4me1-trained classifier. (i) Left: Differentially H3K9me2-bound genes in cluster 4, shown in **g** (two-sided Wilcoxon signed-rank test, log2FC > 0.6, adjusted p-value < 0.01), were tested against MSigDB Hallmark and Breast-related signatures. Shown are the top 5 terms significantly enriched (nominal p-value < 0.05) for genes with upregulation of H3K9me2 on their promoters. Right: UMAP as in **g** colored according to UCell gene-signature scores for the indicated gene sets. Cluster 4 shows elevated H3K9me2 occupancy over luminal-associated genes. (j) Read count per cell (nCount) by cluster (panel **g**). Violin plots show the distribution of individual cells, with the median denoted by a white dot and n indicated for each cluster. Cluster 4 shows the highest number of reads per cell, possibly corresponding to higher global levels of this mark in basal-like cells.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Cross-cohort performance of the EPINUC-based classifier for breast cancer detection. (a,b) Classifier performance in an independently processed, batch-corrected cohort of healthy individuals (n = 19) and luminal BC patients (n = 30), using the model trained on the original cohort in Fig. 4f. (**a**) ROC curve (AUC = 0.98; shading, 95% CI). (**b**) Confusion matrix with specificity set to 0.9. (c,d) ROC curve discriminating healthy individuals (n = 32) from luminal BC patients, diagnosed at the SHEBA medical center, Israel (**c**, n = 16, AUC = 0.96) or Kaplan medical center, Israel (**d**, n = 14, AUC = 0.99). Shading indicates the 95% CI. (e) ROC curve discriminating healthy individuals (n = 32) and triple-negative BC patients (n = 22, AUC = 0.98, shading indicates the 95% CI). (f) Healthy (n = 51) and luminal BC samples from three medical centers (n = 137) were pooled and stratified into 75% training cohort and a 25% test cohort. ROC curves are shown for training (left, n = 38 healthy vs. 103 luminal BC samples, AUC = 0.96) and the test cohort (right, n = 13 healthy vs. 34 luminal BC patients, AUC = 0.94). Shading indicates the 95% CI. (g) AUC estimates for discrimination of healthy individuals (n = 51) from luminal A (n = 98), luminal B (n = 31), or all luminal BC patients (n = 137) across the three medical centers. Points and lines indicate AUCs and 95% CIs, respectively.



Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | Epigenetic differences in the plasma of patients who developed recurrent disease. (a-f) Selected EPINUC features in disease-free versus recurrent-disease groups within clinically defined subsets (see Fig. 5a): >50% ER-positive (**a**, n = 39 vs. 28), >50% PR-positive (**b**, n = 33 vs. 18), luminal A BC (**c**, n = 35 vs. 23), T1-T2 staged tumors (**d**, n = 43 vs. 20), grade 1-2 tumors (**e**, n = 36 vs. 18), and no LVI at diagnosis (**f**, n = 34 vs. 14). Two-sided Wilcoxon tests. (g) Distribution of 29 EPINUC features across luminal BC patients stratified by histological grade (grade 1, n = 29; grade 2, n = 67; grade 3, n = 36). Points represent individual patients. Kruskal-Wallis p-values were Benjamini-Hochberg-corrected across the 29 features. (h) Spearman correlations between mean histone PTM levels measured by CyTOF in tumor cells and EPINUC measurements in matched plasma (n = 10). Correlations with r > 0.2 are pink; P < 0.05 is bold (two-sided Spearman correlation test). H3K4me3 and H3K27me3 scatter plots are shown below, with Theil-Sen fits and 95% CIs. (i) Individual PPS analysis distinguishing disease-free (DF) from relapse (R, left column), or between healthy individuals (H) and each of the luminal outcome group. Color indicates PPS, reflecting predictive power of each individual feature. (j) OncotypeDX scores for patients in cohorts MAYO-1 and MAYO-2 with available measurements (disease-free, n = 40; relapse, n = 9, n.s., non-significant, two-sided Wilcoxon test). In a–g,j, boxes show median and 25th–75th percentiles; whiskers extend to the largest/smallest value within 1.5×IQR.

nature portfolio

Corresponding author(s): Guy Ron and Efrat Shema

Last updated by author(s): 2026/09/12

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software were used for data collection
Data analysis	CellProfiler v3.1.9; DIA-NN v1.8.1; cutadapt v4.4; BSMAP v2.90; GATK v4.6.2.0; MOABS v1.3.9.6; ImageJ v1.53; cellranger-arc v2.2.0; cellranger v10.0.0; Trim Galore! v0.6.5; Bismark v0.24.2; Bowtie2 v2.3.5.1; Samtools v1.16.1; DRAGEN TruSight Oncology 500 analysis software v2.6.2; MaxQuant v1.5.2.8; QualBrowser v2.0.7; FlowJo v10.10.0; CyTOF software v7.1.16401; QuPath v6.0; Cellpose-SAM v4.x; clusterProfiler v4.8.3; WGBStools v0.3.0; RLM v1.2.0; ClubCpG v3.0; EDec v0.9; MACS2 v2.2.9.1; Signac v1.16.0; Seurat v5.3.0; UCell v2.13.1; ppscore v1.3.0; XGBoost v3.0.3; SHAP v0.48.0; and Boruta v4.3. Code used for WGBS, snCUT&Tag and EPINUC analyses is available at https://doi.org/10.5281/zenodo.15336152; CyTOF analysis code is available at https://doi.org/10.5281/zenodo.21982487.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Raw CyTOF and EPINUC measurements are available at https://doi.org/10.5281/zenodo.15336152. Due to protection of personal data requirements, the patient-derived sequencing data generated in this study have been deposited in the European Genome-phenome Archive (EGA) under controlled access, accession EGAS50000002070. Access is subject to approval by the Weizmann Breast Cancer Epigenetics Data Access Committee, consistent with participant consent and ethical approvals. The MARS-seq and H3K4me1 CUT&Tag data of sorted cells have been deposited in the Gene Expression Omnibus (GEO) under accession GSE334437. The MS histone PTM data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD057150.
Publicly available data and resources used in this study include the TCGA-BRCA cohort (dbGaP Study Accession phs000178, Firehose stddata__2016_01_28, DOI: 10.7908/C11G0KM9), the Regner et al. breast cancer single-cell multiome dataset (GSE243526), JASPAR transcription-factor binding profiles (DOI: 10.1093/nar/gkab1113) and scRNA-seq data of human normal mammary epithelial cells (GSE180878).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Participants in this study: women (according to biological sex) aged 18-99, diagnosed with breast cancer by biopsy and scheduled to undergo lumpectomy/mastectomy surgery.
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).* *Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)* *Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	Participants in this study: women (according to biological sex) aged 18-99, diagnosed with breast cancer by biopsy and scheduled to undergo lumpectomy/mastectomy surgery.
Recruitment	Method of recruitment of participants: the attending physicians located women who meet the research criteria and asked for their consent to the researchers' application. The patients participating in the study signed a consent form. Recruitment was based on eligibility and sample availability. As the study required treatment-naïve tumors, the tumor cohort predominantly comprised patients with luminal breast cancer, limiting generalizability to other breast cancer subtypes.
Ethics oversight	The collection and analysis of breast cancer tissues and matching blood samples were approved by the Israeli Ministry of Health Ethics Committees (Helsinki applications KMC-0028-22), followed by approval of the Weizmann Institutional Review Board (IRB, 2314-3). Informed consent was obtained from all individuals. The breast cancer samples for histone PTM MS-based analysis, which comprise 21 tumors, were obtained from patients undergoing surgery at the European Institute of Oncology (Milan) and are part of a study approved by the institute's Ethical Committee with Study UID 2550. Plasma samples and data for luminal breast cancer patients were obtained from Mayo Clinic (IRB approval number 23-011311). Collection of plasma samples from breast cancer patients treated at the Sheba Medical Center was approved by the Institutional Review Board (IRB approval 7509-09) and informed consent was obtained from all patients. Plasma samples from healthy subjects were obtained from the Weizmann Institute of Science (IRB approval number 1850-1). All studies were performed in accordance with the relevant guidelines and regulations, and informed consent was obtained from all participants.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size was determined based on sample availability.
Data exclusions	No data were excluded from the analysis, unless indicated in the figure legends.
Replication	Multiple samples were analyzed as indicated in the figure legends.
Randomization	Randomization was not applicable because participants and biological samples were not assigned to experimental treatment or intervention groups.
Blinding	Blinding was not applied because sample groups were defined by existing clinical information, while quantitative measurements and downstream analyses were performed using standardized workflows.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Catalog numbers for all antibodies are provided in Tables S1 and S5a and in the Methods.
Validation	All primary antibodies are commercially available. For standard applications, manufacturer validation for the relevant species and application was used where available. CyTOF antibodies were quality-controlled for signal specificity and quantitative performance as previously described (Furth et al, 2025; Griess et al, 2025; Harpaz et al, 2022). Cell-identity markers were further evaluated in breast cancer cell lines of known luminal or basal-like identity. EPINUC antibodies were validated as previously described (Fedyuk et al, 2023) and/or in Fig. S6.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	MCF7 and MDA-MB-468 were obtained from the Tissue Culture Cell (TCC) repository at the Weizmann Institute of Science. HCC70, HCC1937 and SUM149 were provided by Prof. Ittai Ben-Porath (HUJI). Single-cell-derived MCF7 clones and matching parental MCF7 were provided by Prof. Uri Ben-David (TAU) or generated in-house by single-cell sorting.
Authentication	The cell lines used in this study were not authenticated.
Mycoplasma contamination	All cell lines tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	None of the cell lines used in this study are listed in the ICLAC Register of Misidentified Cell Lines.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	Ten-week-old female NSG mice were used for implantation of the human breast cancer PDX models TM00112 (PDX-1) and TM00386 (PDX-2).

Wild animals	The study did not involve wild animals.
Reporting on sex	Only female mice were used
Field-collected samples	The study did not involve field-collected samples.
Ethics oversight	All animal studies were approved by the Weizmann Institute's Institutional Animal Care and Use Committee (IACUC) and were performed in accordance with Israeli law.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*