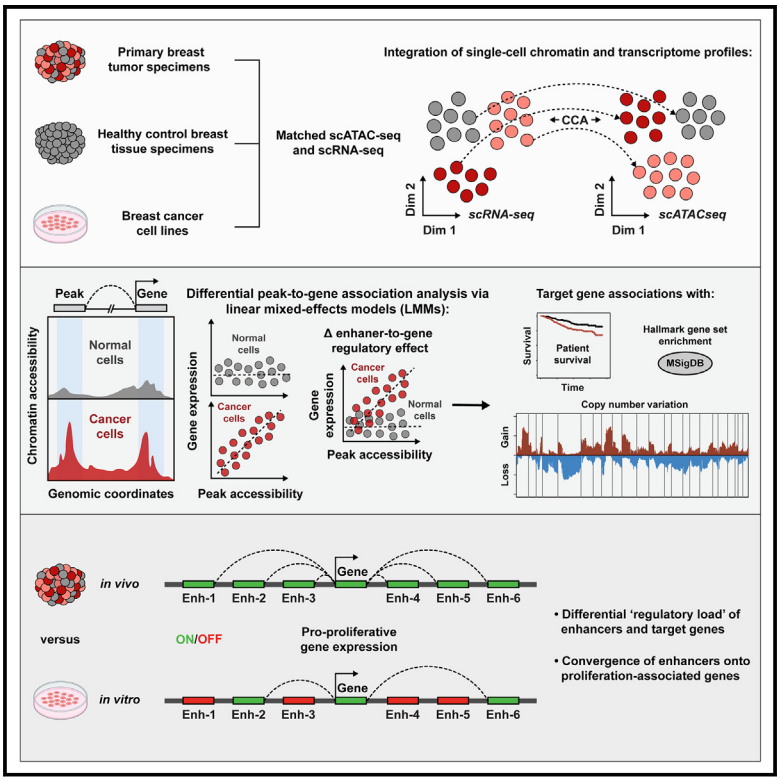
Cell Genomics

Resource

Defining the regulatory logic of breast cancer using single-cell epigenetic and transcriptome profiling

Graphical abstract



Highlights

- Matched scRNA-seq and scATAC-seq of breast tumors, normal tissues, and cell lines
- Linear mixed-effects modeling to annotate cancer-specific enhancers
- Identification of potential silencer-to-enhancer switching events
- Comparison of *in vivo* to *in vitro* enhancer logic in breast cancer cells

Authors

Matthew J. Regner, Susana Garcia-Recio, Aatish Thennavan, ..., Joel S. Parker, Charles M. Perou, Hector L. Franco

Correspondence

hfranco@cccupr.org

In brief

Matched chromatin accessibility and transcriptome profiles at single-cell resolution from breast tumors, normal tissues, and cell lines are used to identify regulatory elements that drive clinically relevant gene expression programs in breast cancer. Regner et al. highlight how cancer cells can rewire regulatory elements to increase the expression of oncogenes.

Regner et al., 2025, Cell Genomics *5*, 100765
February 12, 2025
https://doi.org/10.1016/j.xgen.2025.100765

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Resource

Defining the regulatory logic of breast cancer using single-cell epigenetic and transcriptome profiling

Matthew J. Regner,[1,2] Susana Garcia-Recio,[1,3] Aatish Thennavan,[4] Kamila Wisniewska,[1] Raul Mendez-Giraldez,[1] Brooke Felsheim,[1,2] Philip M. Spanheimer,[1,5] Joel S. Parker,[1,2,3] Charles M. Perou,[1,2,3,6] and Hector L. Franco[1,2,3,7,8,*]

[1]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[2]Curriculum in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[3]Department of Genetics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[4]Department of Systems Biology, UT MD Anderson Cancer Center, Houston, TX 77030, USA
[5]Department of Surgery, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[6]Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[7]Division of Clinical and Translational Cancer Research, University of Puerto Rico Comprehensive Cancer Center, San Juan, PR 00935, USA
[8]Lead contact
*Correspondence: hfranco@cccupr.org
https://doi.org/10.1016/j.xgen.2025.100765

SUMMARY

Annotation of *cis*-regulatory elements that drive transcriptional dysregulation in cancer cells is critical to understanding tumor biology. Herein, we present matched chromatin accessibility (single-cell assay for transposase-accessible chromatin by sequencing [scATAC-seq]) and transcriptome (single-cell RNA sequencing [scRNA-seq]) profiles at single-cell resolution from human breast tumors and healthy mammary tissues processed immediately following surgical resection. We identify the most likely cell of origin for subtype-specific breast tumors and implement linear mixed-effects modeling to quantify associations between regulatory elements and gene expression in malignant versus normal cells. These data unveil cancer-specific regulatory elements and putative silencer-to-enhancer switching events in cells that lead to the upregulation of clinically relevant oncogenes. In addition, we generate matched scATAC-seq and scRNA-seq profiles for breast cancer cell lines, revealing a conserved oncogenic gene expression program between *in vitro* and *in vivo* cells. This work highlights the importance of non-coding regulatory mechanisms that underlie oncogenic processes and the ability of single-cell multi-omics to define the regulatory logic of cancer cells.

INTRODUCTION

Breast cancer (BC) is the most commonly diagnosed cancer among women and accounts for 15% of all female cancer-related deaths in the United States.[1] Treatment strategies and patient prognosis vary by clinical subtype, defined by hormone receptor expression of estrogen receptor (ER) and progesterone receptor (PR) and overexpression and/or amplification of human epidermal growth factor receptor 2 (HER2). BC can also be stratified into five intrinsic molecular subtypes with distinct clinical outcomes: luminal A, luminal B, HER2-enriched, basal-like, and normal-like.[2–5] Together, these form three broad subtypes of BC: luminal ($ER^+/PR^{+/-}$), $HER2^+$ ($HER2^+$, $ER^{+/-}$, $PR^{+/-}$), and triple-negative (ER^-, PR^-, $HER2^-$) BC.[6–8] Several studies have characterized the transcriptional landscapes of these BC subtypes.[9–12] While these studies have been transformative, there has been an increased focus on non-coding regions of the genome, in addition to transcriptomics, for deeper multi-omic insights into BC heterogeneity and its pathogenesis.[13–16]

Non-coding regions contain vast amounts of regulatory information that contribute profoundly to tumor biology.[17,18] Moreover, it has become increasingly evident that regulatory elements (i.e., *cis*-acting enhancers) are rewired in cancer cells to promote growth, survival, and other aggressive phenotypes associated with poor clinical outcome.[19–27] Several studies have used epigenomics, in parallel with transcriptomics, to characterize molecular and clinical heterogeneity of BC revealing extensive variation in enhancer activity across BC subtypes.[13–16] However, most studies to date have done so using bulk genomic sequencing of material collected from heterogeneous mixtures of different cell types, obscuring cancer cell-specific activity of oncogenic enhancers. Therefore, exact mechanisms of gene regulation in the context of BC cells remain elusive.

Breast tumors are complex cellular microenvironments in which various types of malignant and non-malignant cells contribute to a range of clinically relevant biological phenomena, from cancer progression to treatment response.[28–30] It is widely accepted that BC arises from mammary epithelial cells.[31,32] The normal mammary epithelium mainly comprises mature luminal,

Table 1. Abbreviated clinical data and single-cell metadata for each sample

Sample	Type	ER IHC	PR IHC	Her2 IHC	scRNA-seq cells	scATAC-seq cells
Patient 1	normal breast tissue	not performed	not performed	not performed	8,682	8,166
Patient 2	normal breast tissue	not performed	not performed	not performed	6,971	9,337
Patient 3	normal breast tissue	not performed	not performed	not performed	6,368	9,549
Patient 4	normal breast tissue	not performed	not performed	not performed	10,222	6,860
Patient 5	primary breast tumor	–	+	–	6,583	5,647
Patient 6	primary breast tumor	–	–	–	3,965	1,429
Patient 7	primary breast tumor	+	–	–	7,029	4,093
Patient 8	primary breast tumor	+	+	equivocal	10,085	6,490
Patient 9	primary breast tumor	+	+	equivocal	6,907	2,025
Patient 10	primary breast tumor	+	+	equivocal	5,307	5,016
Patient 11	primary breast tumor	+	+	–	2,997	2,743
Patient 12	primary breast tumor	+	+	equivocal	8,490	6,339
Patient 13	primary breast tumor	+	+	–	6,676	3,383
Patient 14-1	primary breast tumor	+	+	–	9,175	8,010
Patient 14-2	primary breast tumor	+	+	–	5,655	4,768
Patient 15	primary breast tumor	+	+	equivocal	6,767	7,193
MCF7	breast cancer cell line	not performed	not performed	not performed	7,331	4,367
T47D	breast cancer cell line	not performed	not performed	not performed	7,796	6,709
HCC1143	breast cancer cell line	not performed	not performed	not performed	6,827	1,313
SUM149PT	breast cancer cell line	not performed	not performed	not performed	8,697	3,949

The last two columns reflect the number of cells obtained post quality control (QC). "Equivocal" in the Her2 IHC (immunohistochemistry) column denotes a Her2 IHC value of 2+. Her2 fluorescence *in situ* hybridization (FISH) results and extended clinical data for each patient sample (de-identified) can be found in Table S1.

luminal progenitor, and basal epithelial cells, all of which have been studied as possible "cell-of-origin" precursors for different BC molecular subtypes.[33–37] Several studies have proposed luminal progenitor and mature luminal cells as likely cell-of-origin precursors for basal-like and luminal BC, respectively.[33–37] However, changes in the gene regulatory landscape between normal mammary epithelial and subtype-specific BC cells are not as well studied, especially at single-cell resolution.

Single-cell genomics has revolutionized our ability to explore cellular heterogeneity of breast tumors, yet most studies have profiled transcriptomes via single-cell RNA sequencing (scRNA-seq).[30,38–44] The single-cell assay for transposase-accessible chromatin by sequencing (scATAC-seq) performs high-throughput profiling of chromatin accessibility, revealing complex facets of gene regulation, including activity of enhancers at single-cell resolution.[45–47] Together, scRNA-seq and scATAC-seq enable linking of regulatory elements to putative target genes, offering key mechanistic insights into the molecular underpinnings of BC by interrogating the regulatory logic of BC cells.[25,37,48–54] We posit that enhancers with increased activity in BC cells, relative to normal mammary epithelial cells, regulate the expression of genes associated with oncogenic processes.

To investigate how regulatory landscapes may become altered in human BCs relative to the normal mammary epithelium, we generated matched scRNA-seq and scATAC-seq profiles for 12 primary breast tumor specimens, four normal breast tissue specimens, and four BC cell lines. These data, encompassing over 200,000 single cells, will serve as an important resource to the single-cell genomics and BC research communities. Moreover, we introduce a novel methodology that implements linear mixed-effects models (LMMs) to quantify associations between regions of chromatin accessibility (i.e., regulatory elements) and gene expression while accounting for important biological and technical variables.[55,56] This approach enabled us to perform differential association analyses between subtype-specific BC cells and their nearest mammary epithelial cells from healthy controls. We also apply the LMM-based method to BC cell lines and compare regulatory landscapes between BC cells *in vivo* and *in vitro*, stratified by molecular subtype. Through these analyses, we identify context-specific mechanisms of gene regulation in BC cells and unveil clinically relevant non-coding mechanisms for BC pathogenesis at single-cell resolution.

RESULTS

Matched scRNA-seq and scATAC-seq of human breast tumors and normal mammary epithelial tissues

Twelve primary breast tumor specimens and four normal mammary tissue specimens were collected from 11 treatment-naive BC patients undergoing surgery with curative intent and four healthy control patients undergoing a reduction mammoplasty procedure, respectively (Tables 1 and S1; Figure 1). Two

CellPress
OPEN ACCESS

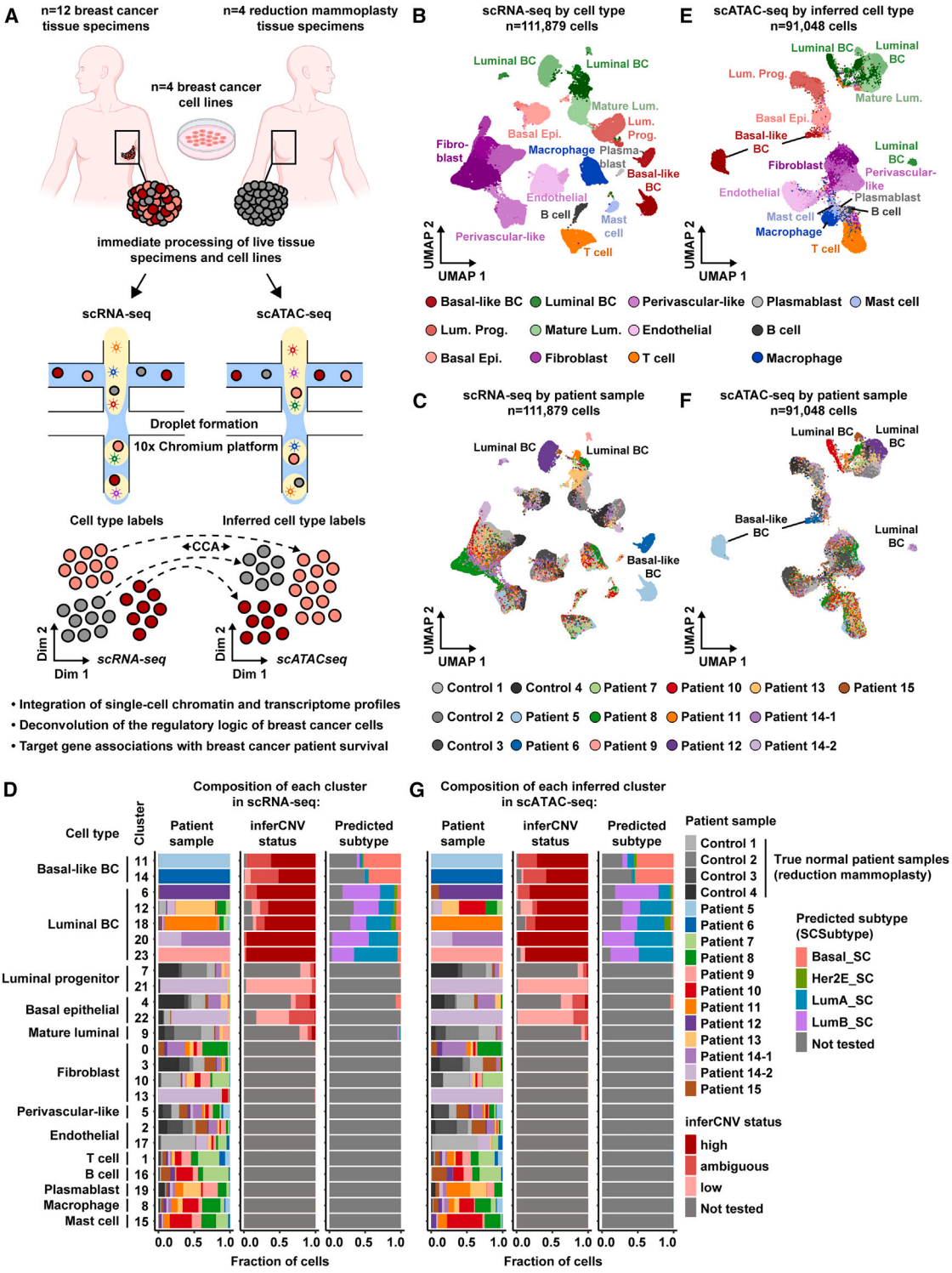


Figure 1. Overview of matched scRNA-seq and scATAC-seq workflow for BC tissue specimens, reduction mammoplasty tissue specimens, and BC cell lines

(A) Schematic diagram of procurement, processing, and downstream analysis of patient samples and cell lines. The female breast and cell line illustrations were created with BioRender.com.

(B) UMAP plot of 111,879 scRNA-seq cells color coded by cell type across 16 patient samples. Color shades denote clusters within each cell type.

(C) UMAP plot of scRNA-seq cells as shown in (B) but color coded by patient sample of origin.

(legend continued on next page)

specimens were collected from BC patient 14 (Tables 1 and S1). Immediately following surgical resection, each specimen was dissociated into a live-cell suspension and prepared for scRNA-seq and scATAC-seq (Figure 1A; STAR Methods). After quality control (QC) for each patient dataset, we obtained 111,879 cells and 91,048 cells profiled by scRNA-seq and scATAC-seq, respectively (Tables 1, S2A, and S3A; Figures S1 and S2).

To analyze scRNA-seq cells from all 16 patient samples, we performed graph-based clustering and visualized all cells in a uniform manifold approximation and projection (UMAP) plot. This showed clusters that were annotated to known cell types (Figures 1B, S3A, and S3B; Table S2A; STAR Methods) and that batch effects were not a major source of variation (Figures 1C and 1D).[30,48,49] To identify malignant BC cells within each patient tumor, we used inferCNV to estimate copy-number variation (CNV) profiles at single-cell resolution as described previously.[30,57–59] Briefly, this procedure involved classifying epithelial cells from each patient tumor into one of three groups (high, ambiguous, or low) based on inferred level of CNV in each cell.[30] Cells classified as inferCNV high were deemed putative cancer cells and were carried forward to molecular subtype prediction.

We predicted the molecular subtype (basal-like, Her2-enriched, luminal A, or luminal B) of each cancer cell within each patient tumor using a previously published method called SCSubtype.[30] These analyses revealed the majority compositions of clusters 11 and 14 as basal-like BC cells from patients 5 and 6, respectively (Figures 1D and S3B; Table S2A). Similarly, the majority of inferCNV-high cells from patients 5 and 6 were predicted to be basal-like BC (Figures S4A and S4B; Table S2A). The majority compositions of clusters 6, 12, 18, 20, and 23 contained mixtures of predicted luminal A and B cells from patients 7–15, suggesting these cells can be referred to hereafter as luminal BC (Figures 1D and S3B; Table S2A). The majority of inferCNV-high cells from patients 7–15 were also predicted to be luminal A or B (Figures S4A and S4B; Table S2A). Finally, we observed that the normal mammary epithelial cell-type clusters of luminal progenitor, basal epithelial, and mature luminal cells were well represented by healthy control samples, indicating a sufficient number of cells for a robust control-group comparison to BC cells (Figures 1D, S3A, and S3B; Table S2A).

To analyze scATAC-seq cells from all 16 patient samples, we performed iterative latent semantic indexing to reduce the dimensionality of the dataset.[46,47,50] Using Seurat's cross-modality integration approach, we then assigned cell-type cluster labels to scATAC-seq cells based on their matching scRNA-seq data (Figures 1E, 1G, S3C, and S3E; Table S3A; STAR Methods).[48–50] This approach also enabled us to assign inferCNV status and predicted subtype to each scATAC-seq cell based on annotations of its nearest neighboring cell in scRNA-seq (Figure 1G; Table S3A; Figures S4C and S4D).[48–50] This showed scATAC-seq cells clustered mainly by cell type, not by patient, consistent with the matching scRNA-seq data (Figures 1F and 1G). In summary, we observed 13 general cell types across the patient dataset, with 24 clusters identified in both scRNA-seq and scATAC-seq. Consistent with previous reports, non-malignant cell-type clusters were well represented across patients, while BC clusters remained highly patient specific, likely due to tumor-unique CNV profiles (Figures 1D and S3B).[25,30,60–62]

To further inform our comparisons of subtype-specific BC cells from primary tumors to normal mammary epithelial cells from healthy controls, we performed an unsupervised clustering analysis of pseudo-bulk transcriptome profiles (Figure S5; STAR Methods). This analysis revealed luminal BC profiles from BC patients clustered with mature luminal profiles from healthy controls, while basal-like profiles from BC patients clustered with luminal progenitor and basal epithelial profiles from healthy controls (Figure S5A). To further investigate this, we performed principal-component analysis (PCA) of the same pseudo-bulk transcriptional profiles, which revealed clear separation of basal-like BC, luminal progenitor, and basal epithelial profiles from luminal BC and mature luminal profiles (Figure S5B). These observations were consistent with previous reports supporting luminal progenitor and mature luminal cells as cell-of-origin precursors to basal-like and luminal BC, respectively.[33,34,37] To this end, we transitioned into subtype-specific analyses of BC (basal-like and luminal BC) compared to their nearest normal mammary epithelial cell types.

Identification of enhancers with cancer-specific regulatory activity in basal-like BC cells

Basal-like BC has been shown to strongly overlap with the triple-negative clinical subtype and portends a poor prognosis, in part due to a lack of targeted therapies.[63–68] To analyze basal-like subtype cells, we merged basal-like BC cells from patients 5 and 6 with luminal progenitor as well as basal epithelial cells from healthy control patients according to the unsupervised pseudo-bulk clustering analysis (Figure S5). This subset resulted in 13,993 cells and 14,038 cells profiled by scRNA-seq and scATAC-seq, respectively (Figure 2A; Tables S2B and S3B). After reclustering scRNA-seq cells and transferring the resulting labels as well as gene expression profiles to scATAC-seq, we found that cells mainly clustered by cell type and not by patient, except for two patient-specific clusters of basal-like BC from patients 5 and 6 (Figure 2B; Tables S2B and S3B).

To interrogate the regulatory logic of basal-like BC cells in comparison to their nearest normal mammary epithelial cells, we first carried out peak calling in scATAC-seq cells to identify putative regulatory elements located in regions of accessible chromatin.[13,50,69,70] We developed a robust LMM-based

(D) Proportion bar charts showing the composition of each cluster in scRNA-seq, in terms of patient sample (left), inferCNV status (middle), and predicted subtype (right). Color-code key is shown to the right.
(E) UMAP plot of 91,048 scATAC-seq cells color coded by inferred cell type across 16 patient samples. Color shades denote clusters within each inferred cell type.
(F) UMAP plot of scATAC-seq cells as shown in (E) but color-coded by patient sample of origin.
(G) Proportion bar charts as in (D) but for the composition of each inferred cluster in scATAC-seq.

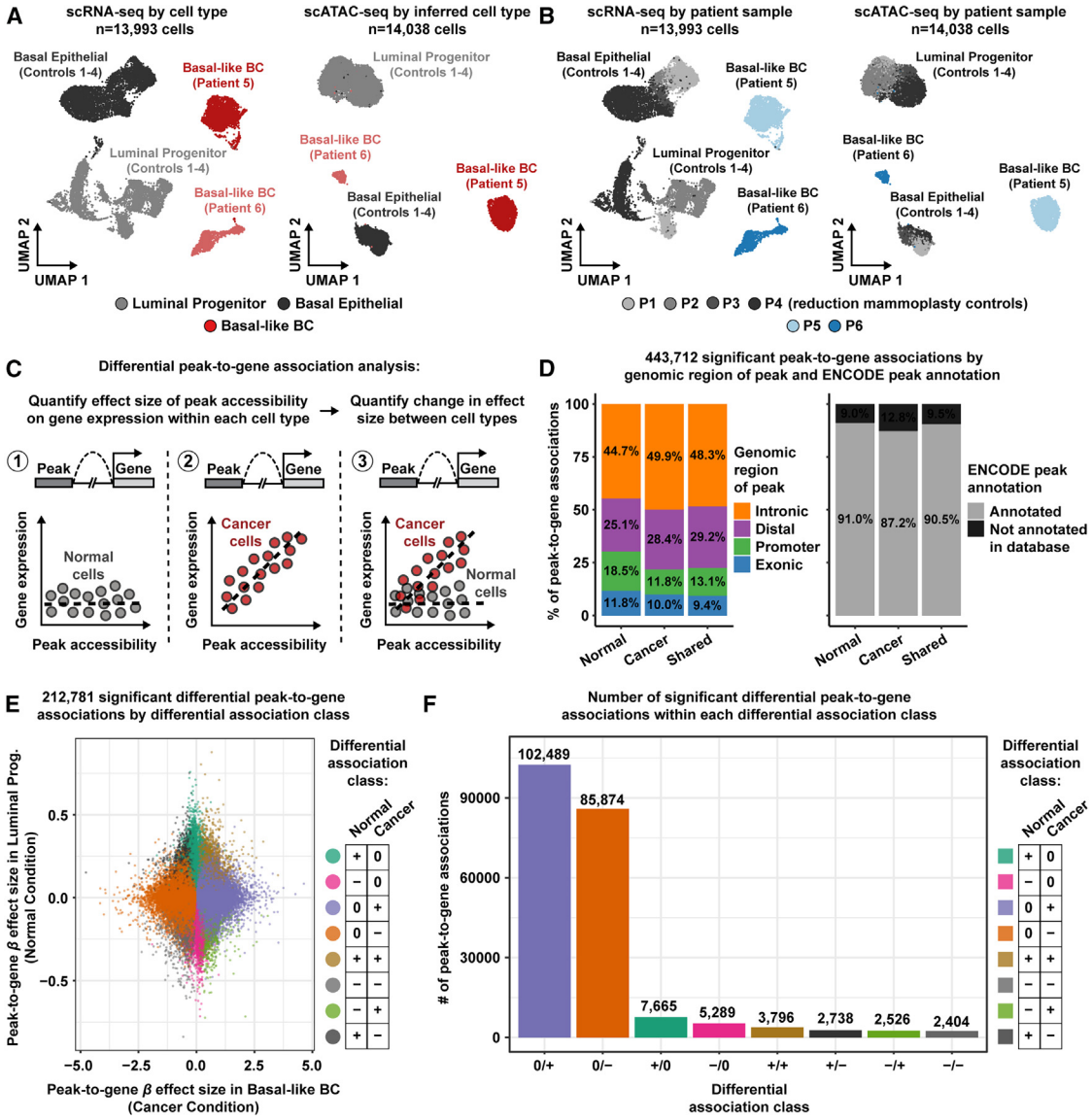


Figure 2. Quantifying the altered regulatory landscape in basal-like BC cells relative to normal luminal progenitor cells

(A) UMAP plot of 13,993 scRNA-seq cells color coded by cell type across six patient samples (left). UMAP plot of 14,038 scATAC-seq cells color coded by inferred cell type across six patient samples (right). Color shades denote clusters within each cell type.

(B) UMAP plots of scRNA-seq cells (left) and scATAC-seq cells (right) as shown in (A) but color coded by patient sample of origin.

(C) Schematic diagram showing the three-step differential peak-to-gene analysis framework.

(D) Proportion bar charts showing the genomic distribution (left) and ENCODE annotation status (right) for 84,975 normal-specific peak-to-gene associations, 337,053 cancer-specific associations, and 21,684 shared associations.

(E) Scatterplot showing effect sizes of significant differential peak-to-gene associations in the cancer condition, comprising basal-like BC cells, and in the normal condition, comprising luminal progenitor cells. Each dot represents a peak-gene pair with a significant change in effect size between conditions and is colored by differential association class (specific change in direction of effect size between conditions).

(F) Bar plot showing the number of differential peak-to-gene associations per differential association class.

strategy to link putative regulatory elements to target genes. This enabled us to quantify the association between peak accessibility and gene expression after accounting for patient. To investigate the regulatory landscapes in basal-like BC cells compared to normal mammary epithelial cells, we performed a two-phased differential peak-to-gene association analysis to link putative regulatory elements to target genes in a context-specific manner (Figure 2C; STAR Methods).

Consistent with previous reports, we focused our analysis on the comparison of basal-like BC cells, referred to hereafter as the "cancer" condition, to luminal progenitor cells, referred to hereafter as "normal."[33,34,37] Using patient-specific metacells

(i.e., aggregates of similar cells), we first quantified, within each condition, the regulatory effect size of peak accessibility on gene expression for every peak within 500 kb of each gene, after accounting for variation between patients (Figure 2C; STAR Methods).[50,55,56] More specifically, for every peak-gene pair tested in each condition, we modeled gene expression as a function of peak accessibility as a fixed effect and patient as a random effect in an LMM. This revealed a total of 443,712 significant peak-to-gene associations (false discovery rate [FDR]-adjusted $p < 1e{-}04$), with 84,975 normal-specific associations, 337,053 cancer-specific associations, and 21,684 shared associations (Figure 2D; Table S4). The majority of these peak-to-gene associations involved peaks located in introns and distal intergenic regions, highlighting the importance of non-coding regulatory information (Figure 2D; Table S4).[25] Moreover, the majority of these peak-to-gene associations also involved peaks annotated by the Encyclopedia of DNA Elements Consortium (ENCODE) database, suggesting they are *bona fide* regulatory elements that provide support for our computational approach (Figure 2D; Table S4).[71,72]

To identify putative regulatory elements with significant differential effects on gene expression between conditions, we combined metacells from both conditions and quantified the change in regulatory effect size between conditions for intronic and distal peak-gene pairs that showed a significant association in at least one condition (Figures 2C and 2D; Table S4).[55,56] We hypothesized that the change in regulatory effect size between conditions could be modeled as an interaction term in the LMM (e.g., modeling gene expression as a function of peak accessibility, condition, and the interaction between condition and accessibility as fixed effects with patient as a random effect). This differential analysis identified 212,781 peak-to-gene associations with significant changes (FDR-adjusted $p < 1e{-}04$) in regulatory effect size between cancer and normal conditions (Figure 2E; Table S4). Of these 212,781 differential peak-to-gene associations, 12,954 differential associations had statistically significant effect sizes in the normal condition but insignificant effect sizes in the cancer condition (Figures 2E and 2F; Table S4). These differential associations may represent putative regulatory element-target gene pairings specific to the normal condition. Conversely, 188,363 differential associations had statistically significant effect sizes in the cancer condition but insignificant effect sizes in the normal condition and thus may reflect cancer-specific putative regulatory element-target gene pairings (Figures 2E and 2F; Table S4). The remaining 11,464 differential peak-to-gene associations had statistically significant effect sizes in both conditions, suggesting potential changes in magnitude and/or direction of regulatory effect these putative regulatory elements exert on target gene expression (Figures 2E and 2F; Table S4).

To this end, we further classified differential peak-to-gene associations based on changes in direction between conditions. Differential peak-to-gene associations with significant effect sizes specific to the cancer condition were either positive (n = 102,489) or negative (n = 85,874), suggesting cancer-specific enhancer or silencer-like regulatory relationships in basal-like BC cells (Figure 2F; Figures S6A and S6B; Table S4). Similarly, differential peak-to-gene associations with significant effect sizes specific to the normal condition were either positive (n = 7,665) or negative (n = 5,289), suggesting enhancer or silencer-like regulatory relationships specific to the normal condition (Figure 2F; Figures S6C and S6D; Table S4). The remaining differential peak-to-gene associations with significant effects in both conditions were either positive in both conditions (n = 3,796) or negative in both conditions (n = 2,404) or showed changes in direction between conditions (n = 5,264) (Figures 2F and S6E–S6H; Table S4). This suggests the potential for regulatory switching events, in which a regulatory element may target the same gene but with opposing regulatory effects depending on cell state. Overall, most differential peak-to-gene associations showed positive regulatory effects that were cancer specific, indicative of putative enhancer-gene regulation specific to the cancer condition (Figure 2F; Table S4).

To characterize putative cancer-specific enhancer-regulated genes in basal-like BC cells, we first screened cancer-specific peak-to-gene associations for those that involve upregulated genes (FDR-adjusted $p < 0.05$ and log2FC ≥ 0.58) in basal-like BC cells (cancer condition) relative to luminal progenitor cells (normal condition) profiled by scRNA-seq.[73–75] This resulted in 7,167 cancer-specific peak-to-gene associations involving upregulated genes, and their effect sizes within each condition were visualized in a heatmap (Figure 3A). We observed that 84.4% of 7,167 cancer-specific peak-to-gene associations involved peaks annotated by ENCODE, suggesting they are *bona fide* enhancers, while the remaining likely represent previously unannotated enhancers (Figure 3A).[71,72] In terms of function, 7,167 cancer-specific peak-to-gene associations involved 829 unique genes that were enriched (FDR-adjusted $p < 0.05$) for the hallmark proliferation-associated gene sets E2F TARGETS, G2M CHECKPOINT, and MITOTIC SPINDLE from the Molecular Signatures Database (MSigDB) (Figure 3B).[76–78] Interestingly, a similar analysis applied to 2,526 putative silencer-to-enhancer switching events revealed 87 unique enhancer-regulated genes upregulated in basal-like BC cells that were also enriched for the same hallmark proliferation-associated gene sets (Figures S7A–S7E; Table S4). To compare to enhancer regulation in normal luminal progenitor cells, we performed the same analysis for 401 putative normal-specific enhancers, revealing 274 unique enhancer-regulated genes upregulated in luminal progenitor cells that were enriched for the hallmark gene sets INTERFERON GAMMA RESPONSE, TNFA SIGNALING VIA NFKB, and ANDROGEN RESPONSE (Figures S7F and S7G; Table S4). To assess which transcription factors (TFs) may bind to these cancer- and normal-specific enhancers, we performed a motif analysis, revealing strong enrichment of Fra2 and Atf3 TF motifs in normal-specific enhancer regions and strong enrichment of Jun-AP1 and CTCF TF motifs in cancer-specific enhancer regions (Figure S8; Tables S5A and S5B). This is consistent with previous reports demonstrating elevated AP1 activity and altered CTCF-dependent topologically associating domains in triple-negative breast cancer (TNBC).[79,80] Together, these observations provide support for the putative cancer-specific enhancers we identified and suggest that the activities of these enhancers, specifically in basal-like BC cells, may play critical roles in upregulating genes involved in proliferation.

To screen for clinically relevant genes regulated by cancer-specific enhancers, we developed a prioritization scheme

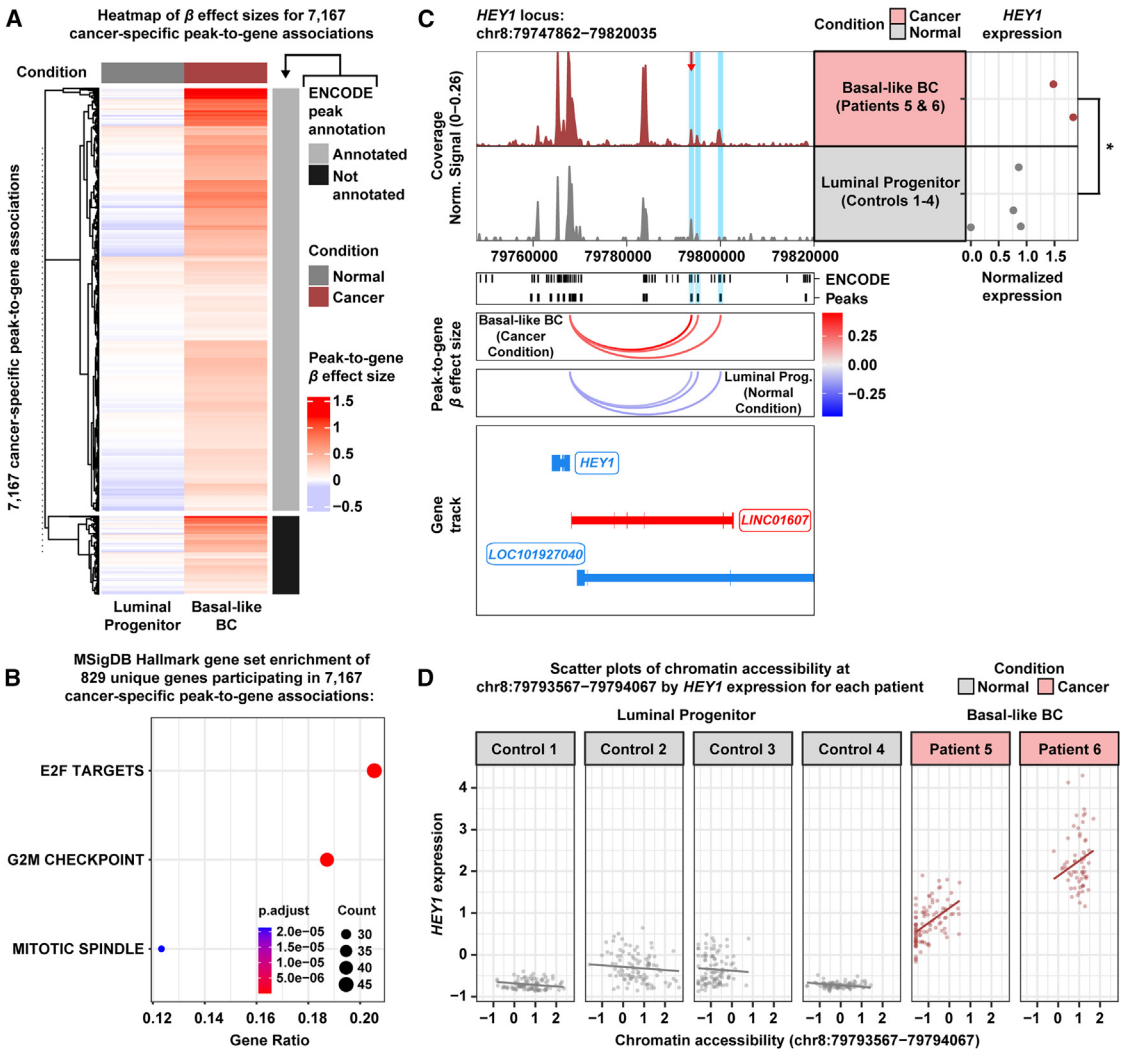


Figure 3. Cancer-specific enhancer regulation of *HEY1* expression in basal-like BC cells

(A) Heatmap of effect sizes for 7,167 cancer-specific peak-to-gene associations in the normal condition, comprising luminal progenitor cells, and in the cancer condition, comprising basal-like BC cells (left). Each row represents a peak-gene pair with a significant change in effect size between conditions. The ENCODE peak annotation column denotes ENCODE annotation status for each cancer-specific peak-to-gene association (right).

(B) Hallmark gene set enrichment analysis of 829 unique genes participating in 7,167 cancer-specific peak-to-gene associations as shown in (A).

(C) Browser track showing the accessibility profile at the *HEY1* locus in cancer (red) and normal (gray) conditions (top left). The putative cancer-specific enhancer with the highest effect size ($\beta = 0.44$, FDR-adjusted $p < 1e{-}04$) on *HEY1* expression is highlighted in light blue and marked by the red arrow. Nearest-neighboring putative cancer-specific enhancers ($\beta = 0.25$ and $\beta = 0.27$, FDR-adjusted $p < 1e{-}04$) are also highlighted in light blue. Matching pseudo-bulk scRNA-seq expression of *HEY1* is shown for each condition (top right). Asterisk denotes a statistically significant difference in gene expression between conditions (FDR-adjusted $p < 0.05$ and log2FC ≥ 0.58). ENCODE regulatory element annotations and peaks called from the scATAC-seq data are shown below the browser track (middle). Peak-to-gene loops show the standardized effect sizes, in each condition, of chromatin accessibility at the putative cancer-specific enhancers on *HEY1* expression (bottom).

(D) Scatterplots of chromatin accessibility at the strongest putative cancer-specific enhancer by the inferred level of *HEY1* expression in scATAC-seq metacells, stratified by patient in the normal (gray) and cancer (red) conditions.

evaluating prognostic status and copy number amplification state of cancer-specific enhancer-regulated genes (Figure S9). For each gene, we performed survival analyses and assessed copy number variation (CNV) status in external patient cohorts (Figures S9 and S10; Table S6). In this prioritization scheme, we first fit Cox proportional hazards models for upregulated genes linked to cancer-specific enhancers using TNBC patients in each dataset (STAR Methods). These analyses revealed that high expression of *HEY1* and *BRSK2* were associated with worse outcomes in three of five TNBC patient datasets (hazard ratio [HR] > 1, Cox $p < 0.01$) (Figures S9 and S10A; Table S6).

We next evaluated frequencies of CNV that affect *HEY1* and *BRSK2* in TCGA TNBC patients. This revealed that >50% of TNBC patients showed copy number gains near the *HEY1* locus

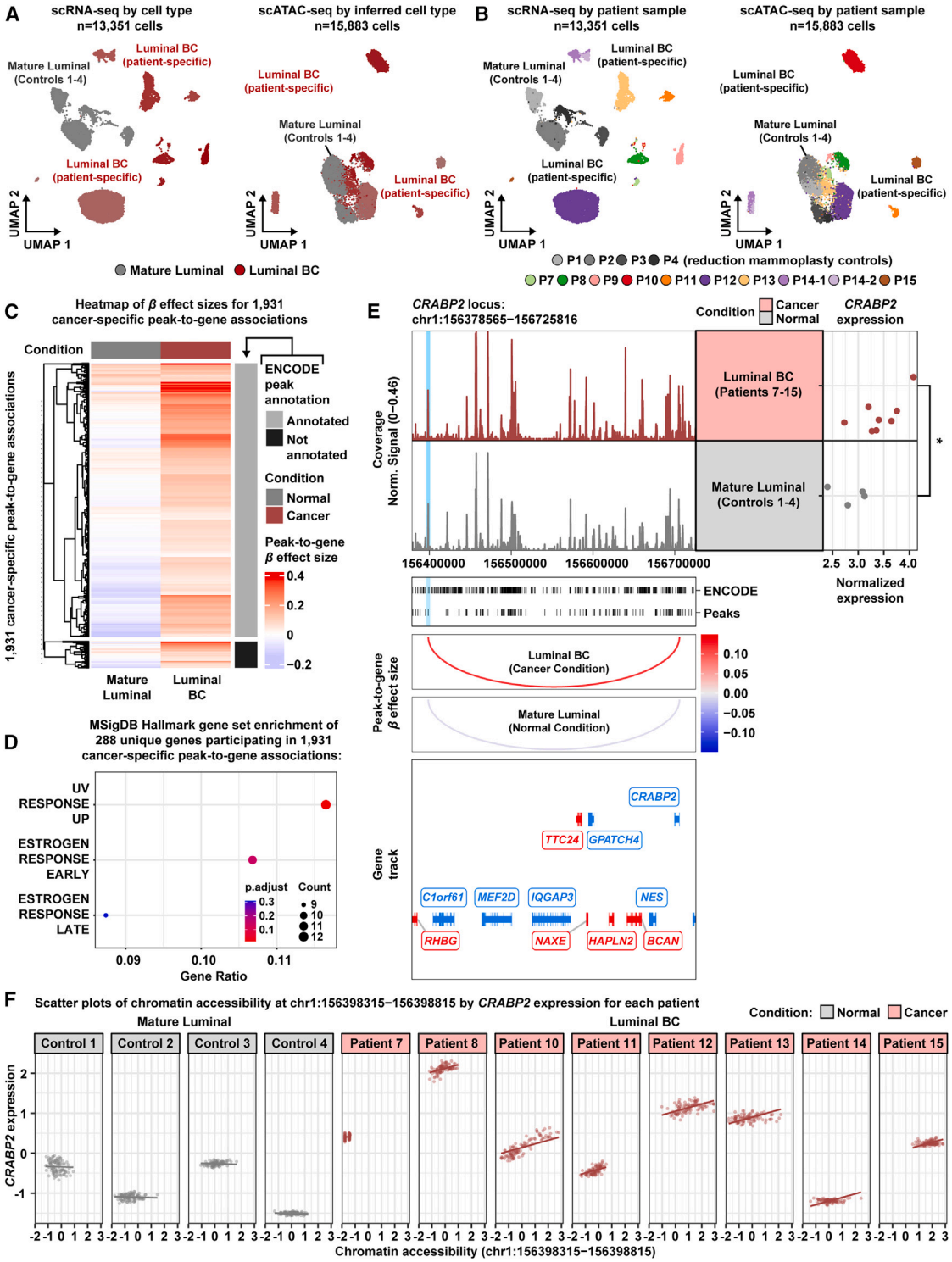


Figure 4. Cancer-specific enhancer regulation of *CRABP2* expression in luminal BC cells

(A) UMAP plot of 13,351 scRNA-seq cells color coded by cell type across 14 patient samples (left). UMAP plot of 15,883 scATAC-seq cells color coded by inferred cell type across 14 patient samples (right). Color shades denote clusters within each cell type.

(B) UMAP plots of scRNA-seq cells (left) and scATAC-seq cells (right) as shown in (A) but color coded by patient sample of origin.

(legend continued on next page)

on chromosome 8, while <10% of TNBC patients showed copy number gains near the *BRSK2* locus on chromosome 11 (Figure S10B). We next investigated the extent to which copy number amplification could explain *HEY1* upregulation relative to enhancer activity. Interestingly, neighboring genes on the same amplicon showed varying levels of expression relative to *HEY1* (Wilcoxon rank-sum tests, $p < 0.05$) (Figure S10C). Moreover, the same neighboring genes showed varying levels of correlation between their expression and copy number state, with *HEY1* itself showing a low correlation (Pearson correlation <0.2) (Figure S10D). These observations suggest *HEY1* upregulation is driven by enhancer activity rather than copy number amplification alone.

We highlight a specific example of cancer-specific enhancer-gene regulation in basal-like BC cells for the upregulated gene *HEY1* (FDR-adjusted $p < 0.05$ and log2FC $\geq$ 0.58), which was linked to six putative cancer-specific enhancers (Figure S11; Table S4).[55,56] *HEY1* is a direct target of the Notch signaling pathway and encodes Hairy/enhancer-of-split related to YRPW motif protein 1, a basic-helix-loop–helix (bHLH) TF.[81,82] Elevated Notch signaling has been observed in a variety of cancers, including basal-like BC.[83–88] The cancer-specific enhancer with the highest regulatory effect size ($\beta = 0.44$, FDR-adjusted $p < 1e{-}04$) on *HEY1* expression was annotated by ENCODE, but did not show a statistically significant increase in chromatin accessibility in basal-like BC cells relative to luminal progenitor cells (Figure 3C).[50,71–73] This suggests the putative regulatory element is accessible in both conditions, but targets *HEY1* only in basal-like BC cells. The nearest neighboring cancer-specific enhancers showed similar regulatory effects on *HEY1* expression ($\beta = 0.25$ and $\beta = 0.27$, FDR-adjusted $p < 1e{-}04$), both of which were annotated by ENCODE, and the second neighboring cancer-specific enhancer had a statistically significant increase in chromatin accessibility in basal-like BC cells (Figure 3C).[50,71–73] To further visualize putative regulatory effects of these cancer-specific enhancers on *HEY1* expression, we plotted the levels of chromatin accessibility at the cancer-specific enhancers by the inferred levels of *HEY1* expression in scATAC-seq metacells from each patient (Figures 3D and S12). This showed that variation in chromatin accessibility at the strongest cancer-specific enhancer was not associated with variation in *HEY1* expression in luminal progenitor cells from healthy controls but was significantly associated with variation in *HEY1* expression in basal-like BC cells from patients 5 and 6 (Figures 3C and 3D). The same was observed for the nearest neighboring cancer-specific enhancers (Figures 3C and S12).

Together, these observations suggest a possible mechanism for *HEY1* upregulation in basal-like BC cells through the activity of tumor unique enhancers that target gene expression in a cancer-specific manner. We note that this example of cancer-specific enhancer-gene regulation is only a glimpse of the altered regulatory landscape in basal-like BC cells, and we have tabulated all putative regulatory element-target gene pairings identified from the basal-like subtype analysis in Table S4, which serves as a resource for future investigations of these regulatory elements.

Cancer-specific regulatory activity of enhancers in luminal BC cells

Luminal BC is often associated with hormone-receptor-positive BC and is the most commonly diagnosed BC among women.[64,89] To analyze cells in the luminal subtype analysis, we merged luminal BC cells from patients 7–15 with mature luminal cells from healthy control patients according to the unsupervised pseudo-bulk clustering analysis (Figure S5). This subset resulted in 13,351 cells and 15,883 cells profiled by scRNA-seq and scATAC-seq, respectively (Figure 4A; Tables S2C and S3C). After reclustering scRNA-seq cells and transferring the resulting labels as well as gene expression profiles to scATAC-seq, we found that cells mainly clustered by patient, consistent with previous reports, and mature luminal cells from healthy controls were represented by a single cluster (Figure 4B; Tables S2C and S3C).[25,30,60–62]

To probe the altered regulatory landscape in luminal BC cells relative to mature luminal cells, we carried out peak calling and performed the two-phased differential peak-to-gene association framework as performed in the basal-like subtype analysis (Figures 2C and S13).[13,50,55,56,69,70] The first phase of the differential peak-to-gene association analysis, quantifying peak-to-gene regulatory effect sizes within each condition independently, yielded results similar to those of the basal-like subtype analysis, with a total of 430,119 significant peak-to-gene associations (FDR-adjusted $p < 1e{-}04$), most of which involved peaks annotated by ENCODE and were located in introns and distal intergenic regions (Figure S13A; Table S7).[25,71,72]

The second phase of the differential peak-to-gene association analysis, quantifying changes in peak-to-gene regulatory effect size between conditions, yielded a total of 135,633 significant differential peak-to-gene associations (FDR-adjusted $p < 1e{-}04$). We note that 5,859 differential associations showed significant

(C) Heatmap of effect sizes for 1,931 cancer-specific peak-to-gene associations in the normal condition, comprising mature luminal cells, and in the cancer condition, comprising luminal BC cells (left). Each row represents a peak-gene pair with a significant change in effect size between conditions. ENCODE peak annotation column denotes ENCODE annotation status for each cancer-specific peak-to-gene association (right).
(D) Hallmark gene set enrichment analysis of 288 unique genes participating in 1,931 cancer-specific peak-to-gene associations as shown in (C).
(E) Browser track showing the accessibility profile at the *CRABP2* locus for the cancer (red) and normal (gray) conditions (top left). The putative cancer-specific enhancer with the highest effect size ($\beta = 0.11$, FDR-adjusted $p < 1e{-}04$) on *CRABP2* expression is highlighted in light blue. Matching pseudo-bulk scRNA-seq expression of *CRABP2* is shown for each condition (top right). Asterisk denotes a statistically significant difference in gene expression between conditions (FDR-adjusted $p < 0.05$ and log2FC $\geq$ 0.58). ENCODE regulatory element annotations and peaks called from the scATAC-seq data are shown below the browser track (middle). Peak-to-gene loops show the standardized effect size, in each condition, of chromatin accessibility at the putative cancer-specific enhancer on *CRABP2* expression (bottom).
(F) Scatterplots of chromatin accessibility at the putative cancer-specific enhancer by the inferred level of *CRABP2* expression in scATAC-seq metacells, stratified by patient in the normal (gray) and cancer (red) conditions.

changes in direction, which again may be interpreted as possible regulatory switching events. However, most differential peak-to-gene associations showed positive effect sizes specific to the cancer condition of luminal BC cells, which again may be interpreted as putative enhancers targeting gene expression in a cancer-specific manner (Figures S13B, S13C, and S14; Table S7).

There were 1,931 cancer-specific peak-to-gene associations involving upregulated genes (FDR-adjusted $p < 0.05$ and log2FC ≥ 0.58), and their effect sizes within each condition were visualized in a heatmap (Figure 4C).[73–75] Of 1,931 cancer-specific peak-to-gene associations, 91.2% involved peaks annotated by ENCODE (Figure 4C).[71,72] To assess function, we observed that the 1,931 cancer-specific peak-to-gene associations involved 288 unique genes that were enriched (FDR-adjusted $p < 0.05$) for the hallmark DNA-damage-associated gene sets UV RESPONSE UP, ESTROGEN RESPONSE EARLY, and ESTROGEN RESPONSE LATE from MSigDB (Figure 4D).[76–78]

We next characterized 2,948 putative silencer-to-enhancer switching events that involved 2,293 unique genes that were enriched for the hallmark gene sets MYC TARGETS V1 and ESTROGEN RESPONSE EARLY (Figures S15A and S15B). There were 67 putative silencer-to-enhancer switching events that involved 56 unique genes upregulated in luminal BC (Figure S15C). The top three most significant hallmark gene sets were ESTROGEN RESPONSE EARLY, ESTROGEN RESPONSE LATE, and INFLAMMATORY RESPONSE (Figures S15D and S15E).

To investigate enhancer regulation in normal mature luminal cells, we performed the same analysis for 350 putative normal-specific enhancers, revealing 122 unique enhancer-regulated genes upregulated in mature luminal cells (Figure S15F). The top three hallmark gene sets were COAGULATION, EPITHELIAL MESENCHYMAL TRANSITION, and HEDGEHOG SIGNALING (Figure S15G).

To investigate TF occupancy at these cancer- and normal-specific enhancers, we carried out a motif analysis, which revealed strong enrichment of BATF and JunB motifs in normal-specific enhancer regions and strong enrichment of CTCF and BORIS motifs in cancer-specific enhancer regions (Figure S16; Tables S5C and S5D). Indeed, CTCF and its paralog BORIS have been shown to play critical roles in estrogen-mediated gene expression in ER+ BC.[90–93] Together, these observations suggest that the activity of these putative cancer-specific enhancers in luminal BC cells may offer mechanistic insights into proliferation and the regulation of DNA-damage-repair pathways in response to estrogen.[94,95]

We next used our prioritization scheme (Figure S9) to screen for clinically relevant genes regulated by putative cancer-specific enhancers in luminal BC cells. High expression of 24 genes was associated with worse outcomes in two of three HR+/HER2− patient datasets (HR > 1, Cox $p < 0.01$) (Figure S17A; Table S6), and five of these genes were affected by copy number amplification events in >50% of TCGA HR+/HER2− patients (Figure S17B). We ranked these five genes by level of expression in scRNA-seq and chose to further investigate one of the most highly expressed genes, *CRABP2* (Figure S9).

To evaluate the extent to which copy number amplification could explain *CRABP2* upregulation, we again analyzed the expression of neighboring genes on the same amplicon and their expression-copy number correlation. These analyses revealed that neighboring genes showed varying levels of expression relative to *CRABP2*, and *CRABP2* showed a low correlation between its expression and its copy number state (Pearson correlation <0.25) (Figures S17C and S17D). Together, these observations suggest that enhancer activity may play an important role in *CRABP2* upregulation independent of copy number amplification events affecting the *CRABP2* locus.

We highlight a specific example of cancer-specific enhancer-gene regulation in luminal BC cells for the upregulated gene *CRABP2* (FDR-adjusted $p < 0.05$ and log2FC ≥ 0.58), which was linked to 14 putative cancer-specific enhancers (Figure S18; Table S7).[55,56] *CRABP2* encodes cellular retinoic acid binding protein 2, which shuttles retinoic acid from the cytosol to the nucleus.[96,97] Interestingly, high expression of *CRABP2* has been reported in a number of cancers, including BC.[98–107] The cancer-specific enhancer with the highest regulatory effect size ($\beta = 0.11$, FDR-adjusted $p < 1e{-}04$) on *CRABP2* expression was annotated by ENCODE but did not show a statistically significant increase in chromatin accessibility in luminal BC cells (Figure 4E).[50,71–73] This suggests the putative regulatory element is accessible in both conditions, but targets *CRABP2* only in luminal BC cells. To visualize this proposed mechanism, we plotted the levels of chromatin accessibility at the cancer-specific enhancer by the inferred levels of *CRABP2* expression in scATAC-seq metacells from each patient (Figure 4F). Variation in chromatin accessibility at this cancer-specific enhancer was not associated with variation in *CRABP2* expression in mature luminal cells from healthy controls but was significantly associated with variation in *CRABP2* expression in luminal BC cells from each BC patient.

Together, these observations describe a potential mechanism for *CRABP2* upregulation in luminal BC cells. We note that this specific example of cancer-specific enhancer-gene regulation is only a snapshot of the altered regulatory landscape in luminal BC cells, and the remaining putative regulatory element-target gene associations are tabulated in Table S7.

Annotation of the enhancer regulatory landscapes in subtype-specific BC cells *in vitro*

To investigate enhancer-regulated gene expression in subtype-specific BC cells *in vitro*, we also generated matched scRNA-seq and scATAC-seq profiles for the basal-like BC cell lines HCC1143 and SUM149PT and luminal BC cell lines MCF7 and T47D (Figure 1A; Table 1).[108] We carried out QC, dimensionality reduction, and cross-modality integration in the cell-line dataset as performed in the patient datasets. This resulted in 30,651 cells and 16,338 cells profiled by scRNA-seq and scATAC-seq, respectively (Table 1; Tables S2D and S3D; Figures 5A and 5B; Figures S19 and S20).

To link putative regulatory elements to target genes in subtype-specific BC cells *in vitro*, we carried out peak calling in scATAC-seq cells and quantified peak-to-gene regulatory effect sizes, using our robust LMM-based approach, within each subtype independently.[13,50,55,56,69,70] This revealed 144,998 significant peak-to-gene associations, with 105,884 associations specific to basal-like BC cells *in vitro*, 30,998 associations

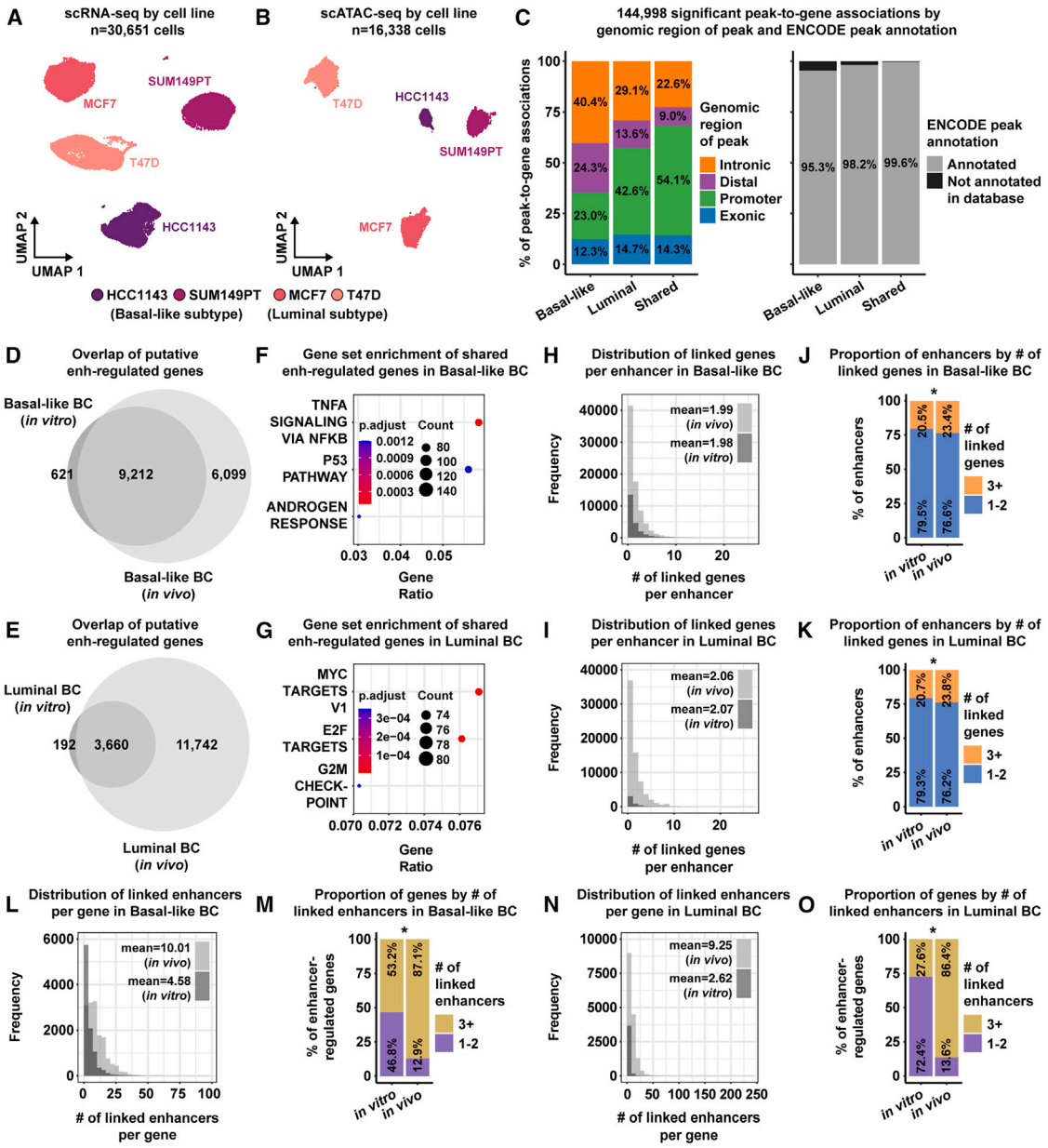


Figure 5. Comparison of enhancer regulatory landscapes between *in vitro* and *in vivo* subtype-specific BC cells

(A) UMAP plot of 30,651 scRNA-seq cells color coded by cell line across four cell line samples.

(B) UMAP plot of 16,338 scATAC-seq cells color coded by inferred cell line across four cell line samples.

(C) Proportion bar charts showing the genomic distribution (left) and ENCODE annotation status (right) for 105,884 basal-like-specific peak-to-gene associations, 30,998 luminal-specific associations, and 8,116 shared associations.

(D) Venn diagram showing the overlap of putative enhancer-regulated genes between basal-like BC cells *in vitro* and *in vivo*.

(E) Venn diagram showing the overlap of putative enhancer-regulated genes between luminal BC cells *in vitro* and *in vivo*.

(F) Hallmark gene set enrichment analysis of 9,212 shared enhancer-regulated genes between basal-like BC cells *in vitro* and *in vivo*.

(G) Hallmark gene set enrichment analysis of 3,660 shared enhancer-regulated genes between luminal BC cells *in vitro* and *in vivo*.

(H) Histograms showing the distributions of linked genes per enhancer for basal-like BC cells *in vitro* and *in vivo*.

(I) Histograms as in (H) but for luminal BC cells *in vitro* and *in vivo*.

(J) Proportion bar charts showing the proportions of enhancers by number of linked genes for basal-like BC cells *in vitro* and *in vivo*. Asterisk denotes a statistically significant difference in the proportion of enhancers that link to three or more genes between BC cells *in vitro* and *in vivo* ($p < 0.01$, Fisher's exact test).

(K) Proportion charts as in (J) but for luminal BC cells *in vitro* and *in vivo*.

(L) Histograms showing the distributions of linked enhancers per gene for basal-like BC cells *in vitro* and *in vivo*.

(legend continued on next page)

specific to luminal BC cells *in vitro*, and 8,116 shared associations (Figure 5C; Table S8). The majority of basal-like-specific peak-to-gene associations involved peaks located in introns and distal intergenic regions, while the majorities of luminal-specific and shared associations involved peaks located in promoters and exonic regions (Figure 5C; Table S8). Again, a strong majority of these peak-to-gene associations involved peaks annotated by ENCODE (Figure 5C; Table S8).[71,72]

We next sought to compare enhancer regulatory landscapes between BC cells *in vitro* and *in vivo* stratified by molecular subtypes (e.g., comparing *in vitro* basal-like BC cells with *in vivo* basal-like BC cells). To this end, we screened peak-to-gene associations identified in subtype-specific BC cells *in vitro* for those with positive effect sizes. The same screening was done for peak-to-gene associations identified *in vivo* from the subtype-specific patient analyses (Figures 2C and 2D; Figure S13A). We then performed overlap analyses of putative enhancer-regulated genes between *in vitro* and *in vivo* subtype-specific BC cells (Figures 5D and 5E; STAR Methods). Of the enhancer-regulated genes *in vitro*, 94% and 95% were also enhancer-regulated *in vivo* for basal-like and luminal BC cells, respectively (Figures 5D and 5E). These shared enhancer-regulated genes in basal-like BC cells were enriched (FDR-adjusted $p < 0.05$) for the hallmark signaling-associated gene sets TNFA SIGNALING VIA NFKB and ANDROGEN RESPONSE as well as the hallmark proliferation-associated gene set P53 PATHWAY from MSigDB (Figure 5F). In luminal BC cells, shared enhancer-regulated genes were enriched (FDR-adjusted $p < 0.05$) for the hallmark proliferation-associated gene sets MYC TARGETS V1, E2F TARGETS, and G2M CHECKPOINT (Figure 5G).

To quantify the "regulatory load" of putative enhancers identified in each setting, we visualized the distributions of linked genes per enhancer (Figures 5H and 5I). In basal-like BC cells, the mean numbers of linked genes per enhancer *in vitro* and *in vivo* were 1.98 and 1.99, respectively (Figure 5H). In luminal BC cells, the mean numbers of linked genes per enhancer *in vitro* and *in vivo* were 2.07 and 2.06, respectively (Figure 5I). We also observed that 23.4% of enhancers identified in basal-like BC cells *in vivo* were linked to three or more genes, compared to only 20.5% of enhancers identified *in vitro* (odds ratio [OR] = 1.19, $p < 0.01$, Fisher's exact test) (Figure 5J). In luminal BC cells, 23.8% of enhancers identified *in vivo* were linked to three or more genes, compared to only 20.7% of enhancers identified *in vitro* (OR = 1.2, $p < 0.01$, Fisher's exact test) (Figure 5K).

The same analyses were performed for the number of linked enhancers per gene (Figures 5L–5O). In basal-like BC cells, the mean numbers of linked enhancers per gene *in vitro* and *in vivo* were 4.58 and 10.01, respectively (Figure 5L). Of enhancer-regulated genes in basal-like BC cells *in vivo*, 87.1% were linked to three or more enhancers, compared to only 53.2% of enhancer-regulated genes *in vitro* (OR = 5.92, $p < 0.01$, Fisher's exact test) (Figure 5M). Similarly, the mean numbers of linked enhancers per gene in luminal BC cells *in vitro* and *in vivo* were 2.62 and 9.25, respectively (Figure 5N). Of enhancer-regulated genes in luminal BC cells *in vivo*, 86.4% were linked to three or more enhancers, compared to only 27.6% of enhancer-regulated genes *in vitro* (OR = 16.69, $p < 0.01$, Fisher's exact test) (Figure 5O).

Overall, these observations suggest that enhancers *in vivo* may regulate more genes compared to enhancers *in vitro*, and genes expressed *in vivo* may be regulated by more enhancers compared to genes expressed *in vitro*. This is possibly due to clonal heterogeneity and/or tumor microenvironment factors that BC cells experience *in vivo* relative to BC cells *in vitro*, which may show less variation in chromatin accessibility and/or gene expression due to the inherent homogeneity of cell lines.

DISCUSSION

BC is the most commonly diagnosed cancer among women and accounts for a significant proportion of female cancer-related deaths, highlighting the need for deeper insights into the molecular underpinnings of BCs that may lead to improved targeted therapies.[1] The compendium presented herein represents a valuable multi-omic resource that unveils transcriptional and regulatory landscapes of human breast tumors and normal mammary epithelial tissues at single-cell resolution. More specifically, our work elucidates transcriptional and regulatory features that distinguish BC cells from their nearest normal precursor cell types by identifying putative enhancers that regulate clinically relevant oncogenic expression programs in a cancer-specific manner (Figures 2, 3, 4, and S5–S18; Tables S4, S5, S6, and S7).[33,34,37,55,56,73,76–78,109–114] These data also enabled us to study transcriptional and regulatory differences between BC cells *in vitro* and *in vivo* (Figure 5; Table S8).[55,56,76–78]

We reiterate three important themes from analyzing these single-cell data. First, we demonstrated how our computational approach for linking putative regulatory elements to target genes accounts for important biological and technical variables when quantifying associations between chromatin accessibility and gene expression (Figures 2, 3, 4, 5, and S13; Tables S4, S7, and S8).[55,56] It has become widely accepted that the activity of *cis*-regulatory elements is highly cell-type specific; therefore, it is critical to stratify by cell type when quantifying peak-to-gene associations from single-cell multi-omic data (Figure 2C).[45–47] Accounting for potential batch and/or sample-specific effects is also important to ensure that technical variation in chromatin accessibility and/or gene expression is not confused with biological variation relevant to the hypothesis.[115,116]

Our computational approach also provides a measure of statistical significance for changes in peak-to-gene regulatory effect size between conditions and/or cell types.[55,56] This allows for inferences about possible changes in magnitude and/or direction of regulatory effect that a regulatory element exerts on target gene expression. Our differential peak-to-gene association analysis allowed us to classify peak-to-gene associations based on changes

(M) Proportion bar charts showing the proportions of genes by number of linked enhancers for basal-like BC cells *in vitro* and *in vivo*. Asterisk denotes a statistically significant difference in the proportion of genes that link to three or more enhancers between BC cells *in vitro* and *in vivo* ($p < 0.01$, Fisher's exact test).
(N) Histograms as in (L) but for luminal BC cells *in vitro* and *in vivo*.
(O) Proportion bar charts as in (M) but for luminal BC cells *in vitro* and *in vivo*.

in direction of effect size between conditions. For both basal-like and luminal subtype analyses, this revealed thousands of cancer-specific associations with positive effect sizes indicative of context-specific putative enhancer activity and evidence to suggest the potential for regulatory switching events that were previously hidden using current peak-to-gene association methods (Figures 2E, 2F, S13B, and S13C).[50,117–119] While similar regulatory mechanisms have been described previously, we highlight that this is one of the first reports of silencer-to-enhancer switching and vice versa in human BC.[120–126]

Next, we highlight that the putative cancer-specific enhancers identified in basal-like and luminal BC cells were linked to genes involved in known oncogenic processes, including proliferation and DNA damage, respectively (Figures 3B and 4D).[76–78] Moreover, we were able to show specific examples of cancer-specific enhancer regulation that not only may be associated with clinical outcomes but also may be amplified through copy number alterations (Figures S10 and S17; Table S6).[109–114] Together, these observations further underscore the importance of non-coding regulatory mechanisms for transcriptional dysregulation in BC cells.

Finally, in the comparisons of subtype-specific BC cells *in vitro* and *in vivo*, our analyses point to a conserved set of genes expressed in both settings for each subtype. We note these conserved sets of expressed genes are associated with known oncogenic processes relevant to BC cells both *in vitro* and *in vivo*, including TNF-alpha signaling and proliferation (Figures 5D–5G).[76–78,127–130] We speculate that perhaps enhancer rewiring or hijacking may be a stochastic process that gets selected for when a specific set of genes favorable to cancer cells becomes upregulated. In summary, this Resource demonstrates important principles of enhancer-gene regulation in BC cells profiled by single-cell multi-omics and serves as an important resource to the field.

Limitations of the study

We acknowledge there are some limitations to our study. First, our study had a limited sample size, especially for patient tumors of the TNBC clinical and/or basal-like molecular subtypes, which could affect the generalizability of the Resource and our observations. However, our study focused on treatment-naive breast tumors, which are difficult to procure as the standard of care shifts toward neoadjuvant treatment prior to surgery.[131–134] Second, libraries for scRNA-seq and scATAC-seq were derived from separate, albeit homogeneous, aliquots of cell suspensions for each patient specimen or cell line. This experimental design requires downstream use of statistical tools for cross-modality integration of these single-cell data, unlike recent methods for profiling the transcriptional and chromatin landscape within the same cell.[37,117] However, cell recovery rate and sequencing depth of these "same cell" protocols can be lower compared to scRNA-seq and scATAC-seq, which can affect the accuracy of downstream analyses.[135] We were also able to validate the performance of cross-modality integration used in our study by leveraging ground-truth cell identities of scATAC-seq cells in the cell-line data. Third, we recognize our survival analyses involved gene expression measurements derived from bulk tissues in contrast to these single-cell data. Finally, we realize our comparisons of BC cells from BC patients to normal mammary epithelial cells from healthy controls may be limited by possible confounding biological factors such as age and menopause status. However, we note our normal control cells from healthy controls represent a true baseline for gene regulation in mammary epithelial tissues, unlike tumor-adjacent normal tissues that may be affected by tumor microenvironment factors and genomic alterations.[136–138] The Resource and our analyses described herein provide an unobscured view of gene regulation in BC cells relative to mammary epithelial cells, highlighting potential avenues for therapeutic interventions.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to the lead contact, Hector L. Franco (hfranco@cccupr.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Processed scRNA-seq data and scATAC-seq data have been deposited at GEO (https://www.ncbi.nlm.nih.gov/geo/) under accession no. GEO: GSE243526. Raw data (10× FASTQs) are available with controlled access via dbGaP under accession no. dbGaP: phs003253.v1.p1 (https://www.ncbi.nlm.nih.gov/gap/).

All code used for the presented analyses is publicly available at the GitHub repository: https://github.com/RegnerM2015/scBreast_scRNA_scATAC_2024. Any additional information required to reanalyze the data reported in this paper is available from the lead contact (hfranco@cccupr.org).

ACKNOWLEDGMENTS

We thank the patients and their families for their generous donations. We thank the University of North Carolina (UNC) Tissue Procurement Facility and UNC Translational Genomics Core Facility for helping us with procuring patient specimens and sequencing genomic libraries. We thank Michele Hayward, Stephanie Metzen, and Matt Soloway at the Office of Genomics Research for help in navigating the institutional review board (IRB) protocols and data submission process to dbGaP and GEO. Finally, we thank members of the Franco and Perou labs for their helpful comments, suggestions, and discussions. This work was supported by grants from the NIH/National Cancer Institute (R01CA273444-03), the Susan G. Komen Breast Cancer Research Foundation (CCR19608601), and the Department of Defense CDMRP Breast Cancer Research Program (BC180450) to H.L.F. Additional support was provided by the UNC Breast Cancer SPORE program (5-P50-CA058223-25) to H.L.F and C.M.P. P.M.S. is supported by National Institutes of Health grant K08CA280388.

AUTHOR CONTRIBUTIONS

H.L.F. and C.M.P. conceived and supervised the study. Patient enrollment and specimen procurement was led by P.M.S. K.W., S.G.-R., and A.T. carried out tissue specimen collection, developed the tissue dissociation protocol, and generated scRNA-seq and scATAC-seq libraries. M.J.R. designed and performed the computational analyses with input from A.T., S.G.-R., R.M.-G., J.S.P., C.M.P., and H.L.F. B.F. generated the Kaplan-Meier survival curves and copy number plots. The manuscript was written by M.J.R. and H.L.F. with input from all authors.

DECLARATION OF INTERESTS

C.M.P is an equity stockholder and consultant of BioClassifier LLC; C.M.P is also listed as an inventor on patent applications for the Breast PAM50 Subtyping assay.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Human patient samples and tissue dissociation
  - Cell culture
- METHOD DETAILS
  - Single-cell sequencing
  - Quantification and quality control (QC) in single-cell RNA-seq
  - Single-cell RNA-seq normalization, feature selection, and clustering
  - Cell type annotation in single-cell RNA-seq
  - Inference of copy number variation (CNV), cancer cell identification, and molecular subtype prediction from single-cell RNA-seq
  - Quality control (QC) in single-cell ATAC-seq
  - Single-cell ATAC-seq quantification, feature selection, and integration with single-cell RNA-seq
  - Unsupervised hierarchical clustering and PCA of pseudo-bulk transcriptomes
  - Differential gene expression and differential peak accessibility testing
  - Peak-to-gene association analysis
  - Overlap analyses of genomic coordinates
  - Gene set enrichment analysis
  - Transcription factor (TF) motif analysis
  - Survival analysis
  - CNV landscape plots
  - Correlation between copy number and expression
  - Prioritization scheme for selection of *HEY1* and *CRABP2*

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2025.100765.

Received: June 13, 2024
Revised: November 4, 2024
Accepted: January 8, 2025
Published: February 5, 2025

REFERENCES

1. Siegel, R.L., Giaquinto, A.N., and Jemal, A. (2024). Cancer statistics, 2024. CA. Cancer J. Clin. *74*, 12–49. https://doi.org/10.3322/caac.21820.
2. Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747–752.
3. Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA *98*, 10869–10874.
4. Sørlie, T., Tibshirani, R., Parker, J., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc. Natl. Acad. Sci. USA *100*, 8418–8423.
5. Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. *27*, 1160–1167.
6. Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., and Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. Am. J. Cancer Res. *5*, 2929–2943.
7. Kim, H.K., Park, K.H., Kim, Y., Park, S.E., Lee, H.S., Lim, S.W., Cho, J.H., Kim, J.Y., Lee, J.E., Ahn, J.S., et al. (2019). Discordance of the PAM50 intrinsic subtypes compared with immunohistochemistry-based surrogate in breast cancer patients: potential implication of genomic alterations of discordance. Cancer Res. Treat. *51*, 737–747.
8. Picornell, A.C., Echavarria, I., Alvarez, E., López-Tarruella, S., Jerez, Y., Hoadley, K., Parker, J.S., del Monte-Millán, M., Ramos-Medina, R., Gayarre, J., et al. (2019). Breast cancer PAM50 signature: correlation and concordance between RNA-Seq and digital multiplexed gene expression technologies in a triple negative breast cancer series. BMC Genom. *20*, 452–511.
9. Edgren, H., Murumagi, A., Kangaspeska, S., Nicorici, D., Hongisto, V., Kleivi, K., Rye, I.H., Nyberg, S., Wolf, M., Borresen-Dale, A.L., and Kallioniemi, O. (2011). Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol. *12*, R6–R13.
10. The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70. https://doi.org/10.1038/nature11412.
11. Shah, S.P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., et al. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature *486*, 395–399. https://doi.org/10.1038/nature10933.
12. Sinicropi, D., Qu, K., Collin, F., Crager, M., Liu, M.L., Pelham, R.J., Pho, M., Dei Rossi, A., Jeong, J., Scott, A., et al. (2012). Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue. PLoS One *7*, e40092.
13. Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W., et al. (2018). The chromatin accessibility landscape of primary human cancers. Science *362*, eaav1898. https://doi.org/10.1126/science.aav1898.
14. Franco, H.L., Nagari, A., Malladi, V.S., Li, W., Xi, Y., Richardson, D., Allton, K.L., Tanaka, K., Li, J., Murakami, S., et al. (02 2018). Enhancer transcription reveals subtype-specific gene expression programs controlling breast cancer pathogenesis. Genome Res. *28*, 159–170. https://doi.org/10.1101/gr.226019.117.
15. Sanghi, A., Gruber, J.J., Metwally, A., Jiang, L., Reynolds, W., Sunwoo, J., Orloff, L., Chang, H.Y., Kasowski, M., and Snyder, M.P. (2021). Chromatin accessibility associates with protein-RNA correlation in human cancer. Nat. Commun. *12*, 5732. https://doi.org/10.1038/s41467-021-25872-1.
16. Huang, H., Hu, J., Maryam, A., Huang, Q., Zhang, Y., Ramakrishnan, S., Li, J., Ma, H., Ma, V.W.S., Cheuk, W., et al. (2021). Defining super-enhancer landscape in triple-negative breast cancer by multiomic profiling. Nat. Commun. *12*, 2242. https://doi.org/10.1038/s41467-021-22445-0.
17. Rheinbay, E., Nielsen, M.M., Abascal, F., Wala, J.A., Shapira, O., Tiao, G., Hornshøj, H., Hess, J.M., Juul, R.I., Lin, Z., et al. (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature *578*, 102–111. https://doi.org/10.1038/s41586-020-1965-x.
18. Zhang, X., and Meyerson, M. (2020). Illuminating the noncoding genome in cancer. Nat. Cancer *1*, 864–872. https://doi.org/10.1038/s43018-020-00114-3.
19. Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., et al. (2013). TERT promoter mutations in familial and sporadic melanoma. Science *339*, 959–961.
20. Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L., and Garraway, L.A. (2013). Highly recurrent TERT promoter mutations in human melanoma. Science *339*, 957–959.
21. Killela, P.J., Reitman, Z.J., Jiao, Y., Bettegowda, C., Agrawal, N., Diaz, L.A., Jr., Friedman, A.H., Friedman, H., Gallia, G.L., Giovanella, B.C., et al. (2013). TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. Proc. Natl. Acad. Sci. USA *110*, 6021–6026.

22. Northcott, P.A., Lee, C., Zichner, T., Stütz, A.M., Erkek, S., Kawauchi, D., Shih, D.J.H., Hovestadt, V., Zapatka, M., Sturm, D., et al. (2014). Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature *511*, 428–434. https://doi.org/10.1038/nature13379.

23. Weischenfeldt, J., Dubash, T., Drainas, A.P., Mardin, B.R., Chen, Y., Stütz, A.M., Waszak, S.M., Bosco, G., Halvorsen, A.R., Raeder, B., et al. (2017). Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. Nat. Genet. *49*, 65–74. https://doi.org/10.1038/ng.3722.

24. Haller, F., Bieg, M., Will, R., Körner, C., Weichenhan, D., Bott, A., Ishaque, N., Lutsik, P., Moskalev, E.A., Mueller, S.K., et al. (2019). Enhancer hijacking activates oncogenic transcription factor NR4A3 in acinic cell carcinomas of the salivary glands. Nat. Commun. *10*, 368. https://doi.org/10.1038/s41467-018-08069-x.

25. Regner, M.J., Wisniewska, K., Garcia-Recio, S., Thennavan, A., Mendez-Giraldez, R., Malladi, V.S., Hawkins, G., Parker, J.S., Perou, C.M., Bae-Jump, V.L., and Franco, H.L. (2021). A multi-omic single-cell landscape of human gynecologic malignancies. Mol. Cell *81*, 4924–4941.e10.

26. Lewis, M.W., Wisniewska, K., King, C.M., Li, S., Coffey, A., Kelly, M.R., Regner, M.J., and Franco, H.L. (2022). Enhancer RNA Transcription Is Essential for a Novel CSF1 Enhancer in Triple-Negative Breast Cancer. Cancers *14*, 1852. https://doi.org/10.3390/cancers14071852.

27. Kelly, M.R., Wisniewska, K., Regner, M.J., Lewis, M.W., Perreault, A.A., Davis, E.S., Phanstiel, D.H., Parker, J.S., and Franco, H.L. (2022). A multi-omic dissection of super-enhancer driven oncogenic gene expression programs in ovarian cancer. Nat. Commun. *13*, 4247. https://doi.org/10.1038/s41467-022-31919-8.

28. Su, S., Chen, J., Yao, H., Liu, J., Yu, S., Lao, L., Wang, M., Luo, M., Xing, Y., Chen, F., et al. (2018). CD10+ GPR77+ cancer-associated fibroblasts promote cancer formation and chemoresistance by sustaining cancer stemness. Cell *172*, 841–856.e16.

29. Cazet, A.S., Hui, M.N., Elsworth, B.L., Wu, S.Z., Roden, D., Chan, C.L., Skhinas, J.N., Collot, R., Yang, J., Harvey, K., et al. (2018). Targeting stromal remodeling and cancer stem cell plasticity overcomes chemoresistance in triple negative breast cancer. Nat. Commun. *9*, 2897. https://doi.org/10.1038/s41467-018-05220-6.

30. Wu, S.Z., Al-Eryani, G., Roden, D.L., Junankar, S., Harvey, K., Andersson, A., Thennavan, A., Wang, C., Torpy, J.R., Bartonicek, N., et al. (2021). 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.

31. Elenbaas, B., Spirio, L., Koerner, F., Fleming, M.D., Zimonjic, D.B., Donaher, J.L., Popescu, N.C., Hahn, W.C., and Weinberg, R.A. (2001). Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. Genes Dev. *15*, 50–65.

32. Keller, P.J., Arendt, L.M., Skibinski, A., Logvinenko, T., Klebba, I., Dong, S., Smith, A.E., Prat, A., Perou, C.M., Gilmore, H., et al. (2012). Defining the cellular precursors to human breast cancer. Proc. Natl. Acad. Sci. USA *109*, 2772–2777.

33. Lim, E., Vaillant, F., Wu, D., Forrest, N.C., Pal, B., Hart, A.H., Asselin-Labat, M.L., Gyorki, D.E., Ward, T., Partanen, A., et al. (2009). Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nat. Med. *15*, 907–913. https://doi.org/10.1038/nm.2000.

34. Molyneux, G., Geyer, F.C., Magnay, F.-A., McCarthy, A., Kendrick, H., Natrajan, R., Mackay, A., Grigoriadis, A., Tutt, A., Ashworth, A., et al. (2010). BRCA1 basal-like breast cancers originate from luminal epithelial progenitors and not from basal stem cells. Cell Stem Cell *7*, 403–417.

35. Nguyen, Q.H., Pervolarakis, N., Blake, K., Ma, D., Davis, R.T., James, N., Phung, A.T., Willey, E., Kumar, R., Jabart, E., et al. (2018). Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. Nat. Commun. *9*, 2028. https://doi.org/10.1038/s41467-018-04334-1.

36. Saeki, K., Chang, G., Kanaya, N., Wu, X., Wang, J., Bernal, L., Ha, D., Neuhausen, S.L., and Chen, S. (2021). Mammary cell gene expression atlas links epithelial cell remodeling events to breast carcinogenesis. Commun. Biol. *4*, 660. https://doi.org/10.1038/s42003-021-02201-2.

37. Terekhanova, N.V., Karpova, A., Liang, W.-W., Strzalkowski, A., Chen, S., Li, Y., Southard-Smith, A.N., Iglesia, M.D., Wendl, M.C., Jayasinghe, R.G., et al. (2023). Epigenetic regulation during cancer transitions across 11 tumour types. Nature *623*, 432–441. https://doi.org/10.1038/s41586-023-06682-5.

38. Azizi, E., Carr, A.J., Plitas, G., Cornish, A.E., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M., et al. (2018). Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell *174*, 1293–1308.e36. https://doi.org/10.1016/j.cell.2018.05.060.

39. Savas, P., Virassamy, B., Ye, C., Salim, A., Mintoff, C.P., Caramia, F., Salgado, R., Byrne, D.J., Teo, Z.L., Dushyanthen, S., et al. (2018). Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. Nat. Med. *24*, 986–993. https://doi.org/10.1038/s41591-018-0078-7.

40. Kim, C., Gao, R., Sei, E., Brandt, R., Hartman, J., Hatschek, T., Crosetto, N., Foukakis, T., and Navin, N.E. (2018). Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. Cell *173*, 879–893.e13. https://doi.org/10.1016/j.cell.2018.03.041.

41. Karaayvaz, M., Cristea, S., Gillespie, S.M., Patel, A.P., Mylvaganam, R., Luo, C.C., Specht, M.C., Bernstein, B.E., Michor, F., and Ellisen, L.W. (2018). Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nat. Commun. *9*, 3588. https://doi.org/10.1038/s41467-018-06052-0.

42. Xu, K., Wang, R., Xie, H., Hu, L., Wang, C., Xu, J., Zhu, C., Liu, Y., Gao, F., Li, X., et al. (2021). Single-cell RNA sequencing reveals cell heterogeneity and transcriptome profile of breast cancer lymph node metastasis. Oncogenesis *10*, 66. https://doi.org/10.1038/s41389-021-00355-6.

43. Pal, B., Chen, Y., Vaillant, F., Capaldo, B.D., Joyce, R., Song, X., Bryant, V.L., Penington, J.S., Di Stefano, L., Tubau Ribera, N., et al. (2021). A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. EMBO J. *40*, e107333. https://doi.org/10.15252/embj.2020107333.

44. Zhang, Y., Zhen, F., Sun, Y., Han, B., Wang, H., Zhang, Y., Zhang, H., and Hu, J. (2023). Single-cell RNA sequencing reveals small extracellular vesicles derived from malignant cells that contribute to angiogenesis in human breast cancers. J. Transl. Med. *21*, 570. https://doi.org/10.1186/s12967-023-04438-3.

45. Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490. https://doi.org/10.1038/nature14590.

46. Cusanovich, D.A., Daza, R., Adey, A., Pliner, H.A., Christiansen, L., Gunderson, K.L., Steemers, F.J., Trapnell, C., and Shendure, J. (2015). Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science *348*, 910–914. https://doi.org/10.1126/science.aab1601.

47. Satpathy, A.T., Granja, J.M., Yost, K.E., Qi, Y., Meschi, F., McDermott, G.P., Olsen, B.N., Mumbach, M.R., Pierce, S.E., Corces, M.R., et al. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat. Biotechnol. *37*, 925–936. https://doi.org/10.1038/s41587-019-0206-z.

48. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

49. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

50. Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., and Greenleaf, W.J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat. Genet. *53*, 403–411. https://doi.org/10.1038/s41588-021-00790-6.

51. Granja, J.M., Klemm, S., McGinnis, L.M., Kathiria, A.S., Mezger, A., Corces, M.R., Parks, B., Gars, E., Liedtke, M., Zheng, G.X.Y., et al. (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat. Biotechnol. *37*, 1458–1465. https://doi.org/10.1038/s41587-019-0332-7.

52. Xu, K., Zhang, W., Wang, C., Hu, L., Wang, R., Wang, C., Tang, L., Zhou, G., Zou, B., Xie, H., et al. (2021). Integrative analyses of scRNA-seq and scATAC-seq reveal CXCL14 as a key regulator of lymph node metastasis in breast cancer. Hum. Mol. Genet. *30*, 370–380. https://doi.org/10.1093/hmg/ddab042.

53. Kumegawa, K., Takahashi, Y., Saeki, S., Yang, L., Nakadai, T., Osako, T., Mori, S., Noda, T., Ohno, S., Ueno, T., and Maruyama, R. (2022). GRHL2 motif is associated with intratumor heterogeneity of cis-regulatory elements in luminal breast cancer. npj Breast Cancer *8*, 70. https://doi.org/10.1038/s41523-022-00438-6.

54. Kim, H., Wisniewska, K., Regner, M.J., Thennavan, A., Spanheimer, P.M., and Franco, H.L. (2023). Single-Cell Transcriptional and Epigenetic Profiles of Male Breast Cancer Nominate Salient Cancer-Specific Enhancers. Int. J. Mol. Sci. *24*, 13053. https://doi.org/10.3390/ijms241713053.

55. Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. Preprint at: arXiv. arXiv:14065823

56. Kuznetsova, A., Brockhoff, P.B., and Christensen, R.H.B. (2017). lmerTest package: tests in linear mixed effects models. J. Stat. Softw. *82*, 1–26.

57. Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189–196.

58. Tickle TI, Georgescu, C., Brown, M. & Haas, B. 2019 inferCNV of the Trinity CTAT Project. https://github.com/broadinstitute/inferCNV

59. Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M., et al. (2019). An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell *178*, 835–849.e21.

60. Izar, B., Tirosh, I., Stover, E.H., Wakiro, I., Cuoco, M.S., Alter, I., Rodman, C., Leeson, R., Su, M.J., Shah, P., et al. (2020). A single-cell landscape of high-grade serous ovarian cancer. Nat. Med. *26*, 1271–1279. https://doi.org/10.1038/s41591-020-0926-0.

61. Lambrechts, D., Wauters, E., Boeckx, B., Aibar, S., Nittner, D., Burton, O., Bassez, A., Decaluwé, H., Pircher, A., Van den Eynde, K., et al. (2018). Phenotype molding of stromal cells in the lung tumor microenvironment. Nat. Med. *24*, 1277–1289. https://doi.org/10.1038/s41591-018-0096-5.

62. Chen, Y.-P., Yin, J.-H., Li, W.-F., Li, H.J., Chen, D.P., Zhang, C.J., Lv, J.W., Wang, Y.Q., Li, X.M., Li, J.Y., et al. (2020). Single-cell transcriptomics reveals regulators underlying immune cell diversity and immune subtypes associated with prognosis in nasopharyngeal carcinoma. Cell Res. *30*, 1024–1042.

63. Bertucci, F., Finetti, P., Cervera, N., Esterni, B., Hermitte, F., Viens, P., and Birnbaum, D. (2008). How basal are triple-negative breast cancers? Int. J. Cancer *123*, 236–240.

64. de Ronde, J.J., Hannemann, J., Halfwerk, H., Mulder, L., Straver, M.E., Vrancken Peeters, M.J.T.F.D., Wesseling, J., van de Vijver, M., Wessels, L.F.A., and Rodenhuis, S. (2010). Concordance of clinical and molecular breast cancer subtyping in the context of preoperative chemotherapy response. Breast Cancer Res. Treat. *119*, 119–126. https://doi.org/10.1007/s10549-009-0499-6.

65. Dogra, A., Mehta, A., and Doval, D.C. (2020). Are basal-like and non-basal-like triple-negative breast cancers really different? J. Oncol. *2020*, 4061063.

66. Shao, F., Sun, H., and Deng, C.-X. (2017). Potential therapeutic targets of triple-negative breast cancer based on its intrinsic subtype. Oncotarget *8*, 73329–73344.

67. Carey, L. (2015). Old drugs, new tricks for triple-negative breast cancer. Lancet Oncol. *16*, 357–359.

68. Abramson, V.G., Lehmann, B.D., Ballinger, T.J., and Pietenpol, J.A. (2015). Subtyping of triple-negative breast cancer: implications for therapy. Cancer *121*, 8–16.

69. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137.

70. Liu, T. (2014). Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells. Methods Mol. Biol. *1150*, 81–95. https://doi.org/10.1007/978-1-4939-0512-6_4.

71. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.

72. ENCODE Project Consortium; Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. https://doi.org/10.1038/s41586-020-2493-4.

73. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

74. Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J.E., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. Nat. Commun. *12*, 5692. https://doi.org/10.1038/s41467-021-25960-2.

75. Murphy, A.E., and Skene, N.G. (2022). A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. Nat. Commun. *13*, 7851. https://doi.org/10.1038/s41467-022-35519-4.

76. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425. https://doi.org/10.1016/j.cels.2015.12.004.

77. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS A J. Integr. Biol. *16*, 284–287.

78. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation *2*, 100141.

79. Chen, H., Padia, R., Li, T., Li, Y., Li, B., Jin, L., and Huang, S. (2021). Signaling of MK2 sustains robust AP1 activity for triple negative breast cancer tumorigenesis through direct phosphorylation of JAB1. npj Breast Cancer *7*, 91. https://doi.org/10.1038/s41523-021-00300-1.

80. Kim, T., Han, S., Chun, Y., Yang, H., Min, H., Jeon, S.Y., Kim, J.I., Moon, H.G., and Lee, D. (2022). Comparative characterization of 3D chromatin organization in triple-negative breast cancers. Exp. Mol. Med. *54*, 585–600. https://doi.org/10.1038/s12276-022-00768-2.

81. Leimeister, C., Externbrink, A., Klamt, B., and Gessler, M. (1999). Hey genes: a novel subfamily of hairy- and Enhancer of split related genes specifically expressed during mouse embryogenesis. Mech. Dev. *85*, 173–177. https://doi.org/10.1016/S0925-4773(99)00080-5.

82. Han, L., Diehl, A., Nguyen, N.K., Korangath, P., Teo, W., Cho, S., Kominsky, S., Huso, D.L., Feigenbaum, L., Rein, A., et al. (2014). The Notch Pathway Inhibits TGFβ Signaling in Breast Cancer through

HEYL-Mediated Crosstalk. Cancer Res. *74*, 6509–6518. https://doi.org/10.1158/0008-5472.CAN-14-0816.

83. Chan, S.M., Weng, A.P., Tibshirani, R., Aster, J.C., and Utz, P.J. (2007). Notch signals positively regulate activity of the mTOR pathway in T-cell acute lymphoblastic leukemia. Blood *110*, 278–286. https://doi.org/10.1182/blood-2006-08-039883.

84. Pinnix, C.C., and Herlyn, M. (2007). The many faces of Notch signaling in skin-derived cells. Pigment Cell Res. *20*, 458–465. https://doi.org/10.1111/j.1600-0749.2007.00410.x.

85. Chen, Y., De Marco, M.A., Graziani, I., Gazdar, A.F., Strack, P.R., Miele, L., and Bocchetta, M. (2007). Oxygen Concentration Determines the Biological Effects of NOTCH-1 Signaling in Adenocarcinoma of the Lung. Cancer Res. *67*, 7954–7959. https://doi.org/10.1158/0008-5472.CAN-07-1229.

86. Reedijk, M., Odorcic, S., Chang, L., Zhang, H., Miller, N., McCready, D.R., Lockwood, G., and Egan, S.E. (2005). High-level Coexpression of JAG1 and NOTCH1 Is Observed in Human Breast Cancer and Is Associated with Poor Overall Survival. Cancer Res. *65*, 8530–8537. https://doi.org/10.1158/0008-5472.CAN-05-1069.

87. Dickson, B.C., Mulligan, A.M., Zhang, H., Lockwood, G., O'Malley, F.P., Egan, S.E., and Reedijk, M. (2007). High-level JAG1 mRNA and protein predict poor outcome in breast cancer. Mod. Pathol. *20*, 685–693. https://doi.org/10.1038/modpathol.3800785.

88. Xu, K., Usary, J., Kousis, P.C., Prat, A., Wang, D.Y., Adams, J.R., Wang, W., Loch, A.J., Deng, T., Zhao, W., et al. (2012). Lunatic Fringe Deficiency Cooperates with the Met/Caveolin Gene Amplicon to Induce Basal-like Breast Cancer. Cancer Cell *21*, 626–641. https://doi.org/10.1016/j.ccr.2012.03.041.

89. Dunnwald, L.K., Rossing, M.A., and Li, C.I. (2007). Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. Breast Cancer Res. *9*, R6. https://doi.org/10.1186/bcr1639.

90. Korkmaz, G., Manber, Z., Lopes, R., Prekovic, S., Schuurman, K., Kim, Y., Teunissen, H., Flach, K., Wit, E.d., Galli, G.G., et al. (2019). A CRISPR-Cas9 screen identifies essential CTCF anchor sites for estrogen receptor-driven breast cancer cell proliferation. Nucleic Acids Res. *47*, 9557–9572. https://doi.org/10.1093/nar/gkz675.

91. Ross-Innes, C.S., Brown, G.D., and Carroll, J.S. (2011). A co-ordinated interaction between CTCF and ER in breast cancer cells. BMC Genom. *12*, 593. https://doi.org/10.1186/1471-2164-12-593.

92. D'Arcy, V., Pore, N., Docquier, F., Abdullaev, Z.K., Chernukhin, I., Kita, G.X., Rai, S., Smart, M., Farrar, D., Pack, S., et al. (2008). BORIS, a paralogue of the transcription factor, CTCF, is aberrantly expressed in breast tumours. Br. J. Cancer *98*, 571–579. https://doi.org/10.1038/sj.bjc.6604181.

93. Akhtar, M.S., Akhter, N., Talat, A., Alharbi, R.A., Sindi, A.A.A., Klufah, F., Alyahyawi, H.E., Alruwetei, A., Ahmad, A., Zamzami, M.A., et al. (2023). Association of mutation and expression of the brother of the regulator of imprinted sites (BORIS) gene with breast cancer progression. Oncotarget *14*, 528–541.

94. Yager, J.D., and Davidson, N.E. (2006). Estrogen carcinogenesis in breast cancer. N. Engl. J. Med. *354*, 270–282.

95. Williamson, L.M., and Lees-Miller, S.P. (2011). Estrogen receptor α-mediated transcription induces cell cycle-dependent DNA double-strand breaks. Carcinogenesis *32*, 279–285. https://doi.org/10.1093/carcin/bgq255.

96. Donovan, M., Olofsson, B., Gustafson, A.-L., Dencker, L., and Eriksson, U. (1995). The cellular retinoic acid binding proteins. J. Steroid Biochem. Mol. Biol. *53*, 459–465.

97. Sessler, R.J., and Noy, N. (2005). A ligand-activated nuclear localization signal in cellular retinoic acid binding protein-II. Mol. Cell *18*, 343–353.

98. Han, S.-S., Kim, W.J., Hong, Y., Hong, S.H., Lee, S.J., Ryu, D.R., Lee, W., Cho, Y.H., Lee, S., Ryu, Y.J., et al. (2014). RNA sequencing identifies novel markers of non-small cell lung cancer. Lung Cancer *84*, 229–235.

99. Wu, J.-I., Lin, Y.-P., Tseng, C.-W., Chen, H.-J., and Wang, L.-H. (2019). Crabp2 Promotes Metastasis of Lung Cancer Cells via HuR and Integrin β1/FAK/ERK Signaling. Sci. Rep. *9*, 845. https://doi.org/10.1038/s41598-018-37443-4.

100. Liu, C.-L., Hsu, Y.-C., Kuo, C.-Y., Jhuang, J.-Y., Li, Y.-S., and Cheng, S.-P. (2022). CRABP2 Is Associated With Thyroid Cancer Recurrence and Promotes Invasion via the Integrin/FAK/AKT Pathway. Endocrinology *163*, bqac171. https://doi.org/10.1210/endocr/bqac171.

101. Gupta, A., Williams, B.R.G., Hanash, S.M., and Rawwas, J. (2006). Cellular Retinoic Acid–Binding Protein II Is a Direct Transcriptional Target of MycN in Neuroblastoma. Cancer Res. *66*, 8100–8108.

102. Egan, D., Moran, B., Wilkinson, M., Pinyol, M., Guerra, E., Gatius, S., Matias-Guiu, X., Kolch, W., le Roux, C.W., and Brennan, D.J. (2022). CRABP2 – A novel biomarker for high-risk endometrial cancer. Gynecol. Oncol. *167*, 314–322. https://doi.org/10.1016/j.ygyno.2022.09.020.

103. Chen, Q., Tan, L., Jin, Z., Liu, Y., and Zhang, Z. (2020). Downregulation of CRABP2 inhibit the tumorigenesis of hepatocellular carcinoma in vivo and in vitro. BioMed Res. Int. *2020*, 3098327.

104. Zeng, S., Xu, Z., Liang, Q., Thakur, A., Liu, Y., Zhou, S., and Yan, Y. (2023). The prognostic gene CRABP2 affects drug sensitivity by regulating docetaxel-induced apoptosis in breast invasive carcinoma: A pan-cancer analysis. Chem. Biol. Interact. *373*, 110372. https://doi.org/10.1016/j.cbi.2023.110372.

105. Zhao, Y., Sun, H., Zheng, J., Shao, C., and Zhang, D. (2021). Identification of predictors based on drug targets highlights accurate treatment of goserelin in breast and prostate cancer. Cell Biosci. *11*, 5–27.

106. Mei, J., Cai, Y., Chen, L., Wu, Y., Liu, J., Qian, Z., Jiang, Y., Zhang, P., Xia, T., Pan, X., and Zhang, Y. (2023). The heterogeneity of tumour immune microenvironment revealing the CRABP2/CD69 signature discriminates distinct clinical outcomes in breast cancer. Br. J. Cancer *129*, 1645–1657.

107. Feng, X., Zhang, M., Wang, B., Zhou, C., Mu, Y., Li, J., Liu, X., Wang, Y., Song, Z., and Liu, P. (2019). CRABP2 regulates invasion and metastasis of breast cancer through hippo pathway dependent on ER status. J. Exp. Clin. Cancer Res. *38*, 361.

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

109. Győrffy, B. (2021). Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. Comput. Struct. Biotechnol. J. *19*, 4101–4109.

110. Jiang, Y.-Z., Ma, D., Suo, C., Shi, J., Xue, M., Hu, X., Xiao, Y., Yu, K.D., Liu, Y.R., Yu, Y., et al. (2019). Genomic and Transcriptomic Landscape of Triple-Negative Breast Cancers: Subtypes and Treatment Strategies. Cancer Cell *35*, 428–440.e5. https://doi.org/10.1016/j.ccell.2019.02.001.

111. Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346–352. https://doi.org/10.1038/nature10983.

112. Staaf, J., Häkkinen, J., Hegardt, C., Saal, L.H., Kimbung, S., Hedenfalk, I., Lien, T., Sørlie, T., Naume, B., Russnes, H., et al. (2022). RNA sequencing-based single sample predictors of molecular subtype and risk of recurrence for clinical assessment of early-stage breast cancer. npj Breast Cancer *8*, 94. https://doi.org/10.1038/s41523-022-00465-3.

113. Ciriello, G., Gatza, M.L., Beck, A.H., Wilkerson, M.D., Rhie, S.K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., et al. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell *163*, 506–519. https://doi.org/10.1016/j.cell.2015.09.033.

114. Xia, Y., Fan, C., Hoadley, K.A., Parker, J.S., and Perou, C.M. (2019). Genetic determinants of the molecular portraits of epithelial cancers. Nat. Commun. *10*, 5666. https://doi.org/10.1038/s41467-019-13588-2.

115. Baek, S., and Lee, I. (2020). Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation. Comput. Struct. Biotechnol. J. *18*, 1429–1439. https://doi.org/10.1016/j.csbj.2020.06.012.

116. Heumos, L., Schaar, A.C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M.D., Strobl, D.C., Henao, J., Curion, F., et al. (2023). Best practices for single-cell analysis across modalities. Nat. Rev. Genet. *24*, 550–572. https://doi.org/10.1038/s41576-023-00586-w.

117. Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., et al. (2020). Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. Cell *183*, 1103–1116.e20. https://doi.org/10.1016/j.cell.2020.09.056.

118. Stuart, T., Srivastava, A., Madad, S., Lareau, C.A., and Satija, R. (2021). Single-cell chromatin state analysis with Signac. Nat. Methods *18*, 1333–1341. https://doi.org/10.1038/s41592-021-01282-5.

119. Fang, R., Preissl, S., Li, Y., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A.K., Zhou, X., Xie, F., et al. (2021). Comprehensive analysis of single cell ATAC-seq data with SnapATAC. Nat. Commun. *12*, 1337. https://doi.org/10.1038/s41467-021-21583-9.

120. Pang, B., van Weerd, J.H., Hamoen, F.L., and Snyder, M.P. (2023). Identification of non-coding silencer elements and their regulation of gene expression. Nat. Rev. Mol. Cell Biol. *24*, 383–395. https://doi.org/10.1038/s41580-022-00549-9.

121. Rosenbauer, F., Owens, B.M., Yu, L., Tumang, J.R., Steidl, U., Kutok, J.L., Clayton, L.K., Wagner, K., Scheller, M., Iwasaki, H., et al. (2006). Lymphoid cell growth and transformation are suppressed by a key regulatory element of the gene encoding PU.1. Nat. Genet. *38*, 27–37. https://doi.org/10.1038/ng1679.

122. Huang, G., Zhang, P., Hirai, H., Elf, S., Yan, X., Chen, Z., Koschmieder, S., Okuno, Y., Dayaram, T., Growney, J.D., et al. (2008). PU.1 is a major downstream target of AML1 (RUNX1) in adult mouse hematopoiesis. Nat. Genet. *40*, 51–60. https://doi.org/10.1038/ng.2007.7.

123. Gisselbrecht, S.S., Palagi, A., Kurland, J.V., Rogers, J.M., Ozadam, H., Zhan, Y., Dekker, J., and Bulyk, M.L. (2020). Transcriptional Silencers in Drosophila Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts. Mol. Cell *77*, 324–337.e8. https://doi.org/10.1016/j.molcel.2019.10.004.

124. Doni Jayavelu, N., Jajodia, A., Mishra, A., and Hawkins, R.D. (2020). Candidate silencer elements for the human and mouse genomes. Nat. Commun. *11*, 1061. https://doi.org/10.1038/s41467-020-14853-5.

125. Ngan, C.Y., Wong, C.H., Tjong, H., Wang, W., Goldfeder, R.L., Choi, C., He, H., Gong, L., Lin, J., Urban, B., et al. (2020). Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. Nat. Genet. *52*, 264–272. https://doi.org/10.1038/s41588-020-0581-x.

126. Huang, Z., Liang, N., Goñi, S., Damdimopoulos, A., Wang, C., Ballaire, R., Jager, J., Niskanen, H., Han, H., Jakobsson, T., et al. (2021). The corepressors GPS2 and SMRT control enhancer and silencer remodeling via eRNA transcription during inflammatory activation of macrophages. Mol. Cell *81*, 953–968.e9. https://doi.org/10.1016/j.molcel.2020.12.040.

127. Franco, H.L., Nagari, A., and Kraus, W.L. (2015). TNFalpha signaling exposes latent estrogen receptor binding sites to alter the breast cancer cell transcriptome. Mol. Cell *58*, 21–34. https://doi.org/10.1016/j.molcel.2015.02.001.

128. Cai, X., Cao, C., Li, J., Chen, F., Zhang, S., Liu, B., Zhang, W., Zhang, X., and Ye, L. (2017). Inflammatory factor TNF-α promotes the growth of breast cancer via the positive feedback loop of TNFR1/NF-κB (and/or p38)/p-STAT3/HBXIP/TNFR1. Oncotarget *8*, 58338–58352.

129. Zhang, Z., Lin, G., Yan, Y., Li, X., Hu, Y., Wang, J., Yin, B., Wu, Y., Li, Z., and Yang, X.P. (2018). Transmembrane TNF-alpha promotes chemoresistance in breast cancer cells. Oncogene *37*, 3456–3470.

130. Narasimhan, H., Ferraro, F., Bleilevens, A., Weiskirchen, R., Stickeler, E., and Maurer, J. (2022). Tumor Necrosis Factor-α (TNFα) Stimulates Triple-Negative Breast Cancer Stem Cells to Promote Intratumoral Invasion and Neovasculogenesis in the Liver of a Xenograft Model. Biology *11*, 1481.

131. Thompson, A.M., and Moulder-Thompson, S.L. (2012). Neoadjuvant treatment of breast cancer. Ann. Oncol. *23*, x231–x236.

132. Burstein, H.J., Curigliano, G., Loibl, S., Dubsky, P., Gnant, M., Poortmans, P., Colleoni, M., Denkert, C., Piccart-Gebhart, M., Regan, M., et al. (2019). Estimating the benefits of therapy for early-stage breast cancer: the St. Gallen International Consensus Guidelines for the primary therapy of early breast cancer 2019. Ann. Oncol. *30*, 1541–1557.

133. Cardoso, F., Kyriakides, S., Ohno, S., Penault-Llorca, F., Poortmans, P., Rubio, I.T., Zackrisson, S., and Senkus, E.; ESMO Guidelines Committee (2019). Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann. Oncol. *30*, 1194–1220.

134. Pusztai, L., Foldi, J., Dhawan, A., DiGiovanna, M.P., and Mamounas, E.P. (2019). Changing frameworks in treatment sequencing of triple-negative and HER2-positive, early-stage breast cancers. Lancet Oncol. *20*, e390–e396.

135. Lee, M.Y.Y., Kaestner, K.H., and Li, M. (2023). Benchmarking algorithms for joint integration of unpaired and paired single-cell RNA-seq and ATAC-seq data. Genome Biol. *24*, 244.

136. Deng, G., Lu, Y., Zlotnikov, G., Thor, A.D., and Smith, H.S. (1996). Loss of heterozygosity in normal tissue adjacent to breast carcinomas. Science *274*, 2057–2059.

137. Widschwendter, M., Berger, J., Daxenbichler, G., Müller-Holzner, E., Widschwendter, A., Mayr, A., Marth, C., and Zeimet, A.G. (1997). Loss of retinoic acid receptor β expression in breast cancer and morphologically normal adjacent tissue but not in the normal breast tissue distant from the cancer. Cancer Res. *57*, 4158–4161.

138. Cho, Y.H., Yazici, H., Wu, H.-C., Terry, M.B., Gonzalez, K., Qu, M., Dalay, N., and Santella, R.M. (2010). Aberrant promoter hypermethylation and genomic hypomethylation in tumor, adjacent normal tissues and blood from breast cancer patients. Anticancer Res. *30*, 2489–2496.

139. McGinnis, C.S., Murrow, L.M., and Gartner, Z.J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. *8*, 329–337.e4. https://doi.org/10.1016/j.cels.2019.03.003.

140. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (New York: Springer-Verlag).

141. Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849. https://doi.org/10.1093/bioinformatics/btw313.

142. R Core Team. R. A language and environment for statistical computing. https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf.

143. Slyper, M., Porter, C.B.M., Ashenberg, O., Waldman, J., Drokhlyansky, E., Wakiro, I., Smillie, C., Smith-Rosario, G., Wu, J., Dionne, D., et al. (2020). A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. Nat. Med. *26*, 792–802. https://doi.org/10.1038/s41591-020-0844-1.

144. McGinnis, C.S., Murrow, L.M., and Gartner, Z.J. (04 2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. *8*, 329–337.e4. https://doi.org/10.1016/j.cels.2019.03.003.

145. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (New York: Springer-Verlag).

146. Liu, S., Thennavan, A., Garay, J.P., Marron, J.S., and Perou, C.M. (2021). MultiK: an automated tool to determine optimal cluster numbers in single-cell RNA sequencing data. Genome Biol. *22*, 232.

147. Maaninka, K., Lappalainen, J., and Kovanen, P.T. (2013). Human mast cells arise from a common circulating progenitor. J. Allergy Clin. Immunol. *132*, 463–469.e3. https://doi.org/10.1016/j.jaci.2013.02.011.

148. Franzén, O., Gan, L.M., and Björkegren, J.L.M. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database *2019*, baz046. https://doi.org/10.1093/database/baz046.

149. Kimes, P.K., Liu, Y., Neil Hayes, D., and Marron, J.S. (2017). Statistical significance for hierarchical clustering. Biometrics *73*, 811–821. https://doi.org/10.1111/biom.12647.

150. Gu, Z. (2022). Complex heatmap visualization. iMeta *1*, e43. https://doi.org/10.1002/imt2.43.

151. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. Roy. Stat. Soc. B *57*, 289–300.

152. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for Computing and Annotating Genomic Ranges. PLoS Comput. Biol. *9*, e1003118. https://doi.org/10.1371/journal.pcbi.1003118.

153. Yu G. enrichplot: Visualization of Functional Enrichment Result. R package version 1.14.2. 2022.

154. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell *38*, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Collagenase/Hyaluronidase	Stemcell Technologies	Cat#07912
Gentle Collagenase/Hyaluronidase	Stemcell Technologies	Cat#07919
Hydrocortisone	Stemcell Technologies	Cat#74144
Dispase	Stemcell Technologies	Cat#07923
DNase I	Stemcell Technologies	Cat#07900
Critical commercial assays		
Chromium Single Cell 3' GEM, Library & Gel Bead Kit v3	10x Genomics	Cat#PN-1000075
Chromium Single Cell ATAC Library & Gel Bead Kit v1	10x Genomics	Cat#PN-1000110
Chromium Chip B Single Cell Kit	10x Genomics	Cat#PN-10000153
Chromium i7 Multiplex Kit	10x Genomics	Cat#PN-120262
Chromium Chip E Single Cell ATAC Kit	10x Genomics	Cat#PN-1000082
Chromium i7 Multiplex Kit N, Set A	10x Genomics	Cat#PN-1000084
Deposited data		
Processed scRNA-seq data	This Paper	GSE243526
Processed scATAC-seq data	This Paper	GSE243526
Raw scRNA-seq data	This Paper	dbGaP: phs003253.v1.p1
Raw scATAC-seq data	This Paper	dbGaP: phs003253.v1.p1
Experimental models: Cell lines		
MCF-7	ATCC	RRID:CVCL_0031
T47D	ATCC	RRID:CVCL_0553
HCC1143	ATCC	RRID:CVCL_1245
SUM149PT	ATCC	RRID:CVCL_3422
Software and algorithms		
R (v4.1.2)	The R Project for Statistical Computing	https://www.r-project.org/
Seurat (v4.1.0)	Hao et al.[49]	https://satijalab.org/seurat/index.html
ArchR (v1.0.1 or v1.0.2)	Granja et al.[50]	https://www.archrproject.com/
DESeq2 (v1.34.0)	Love et al.[73]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
infercnv (v1.10.1)	Tickle et al.[58]	http://www.bioconductor.org/packages/release/bioc/html/infercnv.html
DoubletFinder (v2.0.3)	McGinnis et al.[139]	https://github.com/chris-mcginnis-ucsf/DoubletFinder
ggplot2 (v3.3.6 or v3.4.3)	Wickham[140]	https://cran.r-project.org/web/packages/ggplot2/index.html
ComplexHeatmap (v2.10.0)	Gu et al.[141]	https://jokergoo.github.io/ComplexHeatmap-reference/book/
Cell Ranger (v3.1.0)	10x Genomics	https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/installation
Cell Ranger ATAC (v1.2.0)	10x Genomics	https://support.10xgenomics.com/single-cell-atac/software/pipelines/latest/installation

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Human patient samples and tissue dissociation

Eleven, treatment naïve, breast cancer patients were enrolled in the 2018 Breast SPORE Project 2 study at the UNC Cancer Hospital (IRB Protocol 17-3228) and underwent curative intent surgical resection (Tables 1 and S1). Additionally, four patients were enrolled who were undergoing reduction mammoplasty surgeries in order to collect normal control samples (Tables 1 and S1). After surgical resection, tissue specimens were sectioned by the pathology department and the remaining tissues were de-identified and collected for this study through the University of North Carolina's Tissue Procurement Facility. The tissue specimens were never fixed or frozen and were transported to the lab immediately after surgical resection on ice in media containing DMEM/F12 media (Gibco) + 1% Penicillin/Streptomycin (Corning). Tissues were dissociated as previously reported.[25] In short, before dissociation, tumor and normal samples were weighed. Tissue mass varied between 0.12 g and 11.5 g. Tissue specimens were then minced and digested overnight in DMEM/F12 + 5% FBS, 15mM HEPES (Gibco), 1x Glutamax (Gibco), 1x Collagenase/Hyaluronidase (Stem Cell Technologies, 07912), 1% Penicillin/Streptomycin (Corning), and 0.48 μg/mL Hydrocortisone (Stem Cell Technologies, 74144) on a stir plate at 37°C and 180 rpm. Some tissues were dissociated with Gentle Collagenase/Hyaluronidase (Stem Cell Technologies, 07919) instead of Collagenase/Hyaluronidase. After digestion, cells were washed twice with cold PBS + 2% FBS and 10mM HEPES (PBS-HF) and centrifuged. To remove red blood cells, the cell pellet was treated with cold Ammonium Chloride Solution (Stem Cell Technologies, 07850) and PBS-HF (ratio 1 Ammonium Chloride: 4 PBS-HF), for 1 minute, then centrifuged. The volume of Ammonium Chloride Solution added was determined by the size of the cell pellet and visual assessment of pink or red color of the pellet. Red blood cell removal was repeated a second time if the pellet still exhibited a pink color after the first treatment. Next, cell pellets were resuspended in 0.05% Trypsin-EDTA (Gibco) and the suspension was gently pipetted up and down for 1 min. Trypsin was then inactivated by adding 10mL PBS-HF solution and the suspension was centrifuged. If cell suspensions were still clumpy after trypsin treatment, cells were resuspended with 1-2 mL Dispase (Stem Cell Technologies, 07923) and 200 μL 1mg/mL DNase I (Stem Cell Technologies, 07900) for 1 min, then inactivated with 10 mL PBS-HF. If the Dispase step was not necessary, cells were treated with DNase I during the trypsinization step. Cells were again centrifuged, then washed in PBS-HF and filtered through a 100μm cell strainer and washed again. The cell pellet was resuspended in DMEM/F12 + 5% FBS using a volume based on the final pellet size and filtered using a 40μm cell strainer. Single-cell suspension concentration and cell viability was measured with the Countess II Automated Cell Counter (Thermo Fisher, AMQAX1000). Cell viability varied between 43% and 94% across all samples, with the majority of suspensions having over 70% viability.

Cell culture

Cell lines were obtained from the American Type Culture Collection (ATCC) and maintained in the Franco lab at UNC Chapel Hill. The MCF-7 and T47D cells were grown in Dulbecco's Modified Eagle Medium (DMEM) (Sigma cat. #30-2002) supplemented with 10% fetal bovine serum (FBS) (Sigma) and 1% penicillin/streptomycin (Corning). The HCC1143, and SUM149PT cells were grown in RPMI-1640 media (Sigma) supplemented with 10% FBS (Sigma) and 1% penicillin/streptomycin (Corning). All cell lines were grown adherent in a 5% CO2 incubator set to 37°C with low passage stocks reserved in liquid nitrogen. Cell lines were authenticated by ATCC and tested for mycoplasma prior to use.

METHOD DETAILS

Single-cell sequencing

Cell suspensions were next used for scRNA-seq and scATAC-seq library prep. For scRNA-seq, cell suspensions were diluted to 1200 cells/μL and 10,000 cells were used in library generation with the following 10x Genomics Single Cell 3' kits: Chromium Single Cell 3' GEM, Library & Gel Bead Kit v3 (PN-1000075), Chromium Chip B Single Cell Kit (PN-10000153), and Chromium i7 Multiplex Kit (PN-120262) following the manufacturer's protocol.

For scATAC-seq, 500,000 cells were used in nuclei isolation following the Nuclei Isolation for Single Cell ATAC Sequencing protocol from 10x Genomics. For the lysis step, cells were lysed for 4 min. For the resuspension step, nuclei were resuspended in 50 μL 1x Nuclei Buffer. Nuclei were counted with the Countess II Automated Cell Counter. 10,000 nuclei were used in library preparation using the following 10x Genomics Single Cell ATAC Kits: Chromium Single Cell ATAC Library & Gel Bead Kit v1 (PN-1000110), Chromium Chip E Single Cell ATAC Kit (PN-1000082), and Chromium i7 Multiplex Kit N, Set A (PN-1000084) following the manufacturer's protocol. All libraries were sequenced using the 10X Genomics suggested sequencing parameters on an Illumina NextSeq 500 instrument.

Quantification and quality control (QC) in single-cell RNA-seq

Filtered feature barcode matrices were generated for each patient and cell line sample using Cell Ranger (version 3.1.0) from 10x Genomics. For each sample, the filtered feature barcode matrix was converted into a Seurat object using the *CreateSeuratObject()* function from the Seurat R package.[48,49,142] QC and doublet removal were carried out for each sample individually. Barcodes with at least 500 expressed genes, at least 1,000 UMI counts, and less than 20% mitochondrial counts were deemed high quality cells and were carried forward to doublet detection.[143] Cells predicted as doublets by the DoubletFinder R package were removed from further

downstream analyses.[142,144] After QC and doublet removal for each sample, Seurat's *merge()* function was used to concatenate the individual patient samples and the individual cell line samples, to form the patient cohort dataset and the cell line cohort dataset, respectively (Tables S2A–S2D).[48,49] The distributions of QC metrics, post-QC, for the patient and cell line cohort datasets are visualized in Figures S1A, S1B, S2A, S2B, S19A, S19B, S20A, and S20B.

Single-cell RNA-seq normalization, feature selection, and clustering

All gene expression matrices (for individual samples as well as cohort datasets) were normalized with Seurat's *NormalizeData()* function.[48,49] Seurat's *FindVariableFeatures()* function was used to identify the top 2,000 most variably expressed genes within each individual sample and within each of the cohort datasets (patient cohort, Basal-like subtype cohort, Luminal subtype cohort, and cell line cohort datasets introduced in Figures 1, 2, 4, and 5, respectively).[48,49] The data were scaled with Seurat's *ScaleData()* function using all genes for each individual sample and using only the top variably expressed genes for the remaining cohort datasets.[48,49] For all analyses of each sample individually and each cohort, the top 2,000 most variably expressed genes were used for principal component analysis (PCA) with Seurat's *RunPCA()* function and cells were visualized in a uniform manifold approximation and projection (UMAP) plot using Seurat's *RunUMAP()* function with the first 30 principal components (PCs).[48,49] UMAP plots were then plotted using the ggplot2 R package.[142,145]

For each individual patient sample, cells were clustered using the R package MultiK to identify an optimal number of clusters.[142,146] Note that MultiK applies Seurat's clustering methods over multiple resolution parameters.[48,49,146] Cells within each cohort dataset (patient cohort, Basal-like subtype cohort, Luminal subtype cohort, and cell line cohort) were clustered using Seurat's clustering methods which included building a graph with the *FindNeighbors()* function using the first 30 PCs and identifying clusters with the *FindClusters()* function.[48,49] The resolution parameter in *FindClusters()* was set to 0.4, 0.015, and 0.015 for the patient, Basal-like, and Luminal cohort datasets, respectively.

Cell type annotation in single-cell RNA-seq

Cell type annotation was initially performed within each individual patient sample after pre-processing and clustering with MultiK.[48,49,144,146] For each patient sample, cells were annotated to known cell types using Seurat's canonical correlation analysis (CCA)-based label transfer procedure with a large scRNA-seq breast cancer (BC) reference dataset downloaded from GSE176078.[30,48,49] Clusters within each patient sample were then annotated based on their majority predicted label from the reference-based label transfer procedure. Since mast cells were underrepresented in the reference scRNA-seq dataset, some clusters within each patient dataset were re-annotated to mast cells if the cluster showed significant marker gene expression of *TPSB2* and *TPSAB1* (Bonferroni-corrected p-value < 0.01, log2FC > 0.25).[30,147] Within each patient sample, cluster annotations were verified by visualizing the distributions of cell type gene signature enrichment scores per cluster using Seurat's *AddModuleScore()* function with relevant signatures sourced from PanglaoDB.[48,49,148] The clusters identified in the patient cohort dataset, after combining the individual patient samples, were annotated based on their majority cell type label derived from the annotated clusters in each individual patient sample (Figures 1B–1D, S3A, and S3B; Table S2A).

Inference of copy number variation (CNV), cancer cell identification, and molecular subtype prediction from single-cell RNA-seq

Inferred CNV scores were estimated for individual cells annotated as epithelial within each BC patient sample using the R package inferCNV.[30,57–59] To identify high-confidence cancer cells within each BC patient sample, cells annotated as epithelial were classified into one of three groups: inferCNV high, ambiguous, or inferCNV low, based on the inferred CNV score of each cell as described previously (Figures 1D and S3B).[30,57–59] Cells classified as inferCNV high were deemed putative cancer cells and were carried forward to molecular subtype prediction (Basal, Her2-enriched, Luminal A, or Luminal B) using the SCSubtype method described previously (Figures 1D, S3B, S4A, and S4B; Table S2A).[30] Briefly, this procedure calculates subtype-specific signature enrichment scores for individual cells assigned to one of four molecular subtypes (Basal, Her2-enriched, Luminal A, or Luminal B) based on the highest signature enrichment score for each cell.[30]

Quality control (QC) in single-cell ATAC-seq

A list of unique ATAC-seq fragments with associated barcodes was generated for each patient and cell line sample using Cell Ranger ATAC (version 1.2.0) from 10x Genomics. These lists of unique ATAC-seq fragments per barcode were read into the ArchR R package using the *createArrowFiles()* function to carry out QC and doublet removal for each sample individually.[50,142] Barcodes with at least 1,000 unique fragments, but no more than 100,000, and TSS enrichment scores greater than or equal to 8 were deemed high quality cells and were carried forward to doublet detection.[47] ArchR's *addDoubletScores()* and *filterDoublets()* functions were used to identify and remove cells predicted as doublets from further downstream analyses.[50] The distributions of QC metrics, post-QC, for the patient and cell line cohort datasets are visualized in Figures S1C, S1D, S2C, S2D, S19C, S19D, S20C, and S20D.

Single-cell ATAC-seq quantification, feature selection, and integration with single-cell RNA-seq

To analyze the scATAC-seq cells in the patient cohort analysis, we quantified Tn5 insertion counts in a matrix of contiguous genomic tiles 500 bp in size using ArchR's *addTileMatrix()* function.[50] As described previously, we used the iterative latent semantic indexing

(LSI) procedure implemented in ArchR's *addIterativeLSI()* function to reduce the dimensionality of the dataset.[25,46,47,50] We visualized the cells in a UMAP plot using ArchR's *addUMAP()* function with the top 30 LSI components, as described previously.[25,50] UMAP plots were then plotted using the ggplot2 R package.[142,145] To integrate with matching scRNA-seq cells, we first calculated gene scores for scATAC-seq cells using ArchR's *addGeneScoreMatrix()* function, as described previously.[25,50] Next, we used ArchR's *addGeneIntegrationMatrix()* function to transfer cell type cluster labels and gene expression profiles from scRNA-seq cells to scATAC-seq cells, as described previously (Figures 1E, S3C, and S3E; Table S3A).[25,48–50] Additionally, inferCNV status and predicted subtype labels were assigned to each scATAC-seq cell based on the annotations of its nearest neighboring cell in scRNA-seq (Figures 1G, S4C, S4D; Table S3A) [48–50]. Using the groupList parameter in *addGeneIntegrationMatrix()*, we constrained the integration to cells from the same patient samples to ensure accurate matching of scRNA-seq and scATAC profiles.[50]

The same procedures were applied to analyze the scATAC-seq cells in the Basal-like subtype, Luminal subtype, and cell line cohort analyses, with the exception of using an unconstrained integration in the cell line cohort analysis (i.e., "all versus all") (Figures 2A, 2B, 4A, 4B, 5A, and 5B; Tables S3B–S3D). To evaluate the performance of the label transfer procedure, we leveraged the ground truth identities of the scATAC-seq cells in the cell line cohort analysis (HCC1143, SUM149PT, MCF7, or T47D) to calculate the percentage of scATAC-seq cells correctly assigned to their true cell line identity (99.71%). After transferring labels to scATAC-seq cells in the Basal-like subtype, Luminal subtype, and cell line cohort analyses, peak calling was carried out in each, as described previously, using ArchR's *addGroupCoverages()*, *addReproduciblePeakSet()*, and *addPeakMatrix()* functions.[13,25,50,69,70]

Unsupervised hierarchical clustering and PCA of pseudo-bulk transcriptomes

To perform the pseudo-bulk clustering analysis, we first created pseudo-bulk transcriptomes by summing gene counts across BC cells of the majority subtype within each BC patient sample. This procedure resulted in two Basal-like-specific pseudo-bulk profiles from Patients 5 and 6, and ten Luminal-specific (Luminal A or B) pseudo-bulk profiles from Patients 7-15 (Figure S5). Similarly, we created pseudo-bulk transcriptome profiles for the normal mammary epithelial cell types from healthy controls by summing gene counts across cells of the same cell type within each healthy patient. This procedure resulted in four mature luminal, four basal epithelial, and four luminal progenitor pseudo-bulk profiles all derived from Patients 1-4 (Figure S5).

Only genes expressed across all pseudo-bulk transcriptome profiles were used for downstream analysis to avoid the possible contribution of technical zeros. The resulting matrix of 24 pseudo-bulk profiles was transformed using the regularized logarithm (rlog) transformation from the DESeq2 R package to stabilize variance and account for differences in library size between patient samples.[73,142] After this transformation, the top 10% most variably expressed genes were used for unsupervised hierarchical clustering analysis performed in the SigClust2 R package with Euclidean distance and ward.D2 as the linkage method.[142,149] This resulted in two statistically significant clusters and the dendrogram was further visualized with a heatmap of scaled pseudo-bulk profiles using the ComplexHeatmap R package (Figure S5A).[141,142,150] The same variably expressed genes were used for generating the PCA plots with DESeq2's *plotPCA()* function (Figure S5B).[73]

Differential gene expression and differential peak accessibility testing

We carried out differential gene expression testing after clustering cells within each individual patient sample, using Seurat's *FindAllMarkers()* function with only.pos set to TRUE and test.use set to "wilcox" to identify cluster marker genes.[48,49] Genes with Bonferroni-corrected p-values <= 0.01 were deemed statistically significant marker genes.

For the comparisons of subtype-specific BC cells from BC patients to their nearest normal mammary epithelial cell types from healthy controls, we performed differential gene expression and peak accessibility testing on a pseudo-bulk scale to overcome the pseudo-replication bias in single-cell data.[74,75] Within the Basal-like subtype analysis, we created pseudo-bulk transcriptome profiles by summing gene and peak counts across Basal-like BC cells within each BC patient. The same operation was performed for normal luminal progenitor cells within each healthy patient. This resulted in two Basal-like BC-specific pseudo-bulk profiles from Patients 5 and 6 and four luminal progenitor pseudo-bulk profiles from Patients 1-4. The same procedure was used to construct pseudo-bulk transcriptome profiles within the Luminal subtype analysis, resulting in nine Luminal BC-specific pseudo-bulk profiles and four mature luminal pseudo-bulk profiles. Note that Patient 9 did not have a sufficient number of cells for the peak-to-gene association analysis (n=111) and was therefore excluded from differential gene expression and peak accessibility testing.

Within both Basal-like and Luminal subtype analyses, uninformative genes and peaks with zero counts across all pseudo-bulk profiles were removed. The resulting matrices within the Basal-like and Luminal subtype analyses were read into DESeq2 with the *DESeqDataSetFromMatrix()* function.[73] To verify that potential differences in cell count between pseudo-bulk profiles would not confound differential gene expression and peak accessibility testing, we visualized PCA plots of the pseudo-bulk profiles colored by cell count using DESeq2's *plotPCA()* function within the Basal-like and Luminal subtype analyses.[73] These verification steps confirmed that differences in cell count between pseudo-bulk profiles were not associated with PCs 1 or 2 in both the Basal-like and Luminal subtype analyses.

Pseudo-bulk differential gene expression testing was performed with DESeq2's *DESeq()* function and genes with FDR (Benjamini-Hochberg) adjusted p-values < 0.05 and absolute log2 fold changes > 0.58 were deemed statistically significant differentially expressed genes.[73,151] The same procedure and thresholds were used for pseudo-bulk differential peak accessibility testing to arrive at statistically significant differentially accessible peaks.

Peak-to-gene association analysis

To quantify associations between peak accessibility and gene expression in the Basal-like and Luminal subtype cohorts, we first generated metacells (i.e., aggregates of 100 similar scATAC-seq cells) via a k-nearest neighbor procedure stratified by patient in the low dimensional LSI space for each cohort's scATAC-seq analysis.[50] This procedure resulted in patient-specific metacells that could be classified into the "cancer" or "normal" conditions, based on the sample of origin of constituent scATAC-seq cells within each metacell. Metacells with more than 80% overlap of cell composition with any other metacell from the same patient were removed from further downstream analysis.[50] Peak accessibility and gene expression were summarized for each metacell by summing peak counts and inferred gene expression values (from the matching scRNA-seq data) across scATAC-seq cells within each metacell.[50] The resulting peak and gene expression matrices were normalized to counts per 10,000 and log2-transformed with a pseudo-count of 1.[50] These peak and gene expression matrices were used as input into the peak-to-gene association analyses for each patient cohort dataset. The same set of procedures was applied to the cell line cohort dataset, with the exception of stratifying the k-nearest neighbor procedure by cell line and classifying the resulting cell line-specific metacells into the "Basal-like" or "Luminal" conditions based on the sample of origin of constituent scATAC-seq cells within each metacell.[50]

Within the Basal-like subtype, Luminal subtype, and cell line cohorts, we first performed independent peak-to-gene association analyses in each condition. Using the patient or cell line-specific metacells, we fit a linear mixed-effects model (LMM) with random intercepts, in each condition, using the lmerTest and lme4 R packages, to quantify the effect size of peak accessibility on gene expression for every peak located within 500 kb of each gene.[55,56,142] More specifically, gene expression was modeled as a function of peak accessibility, treated as a fixed effect, and patient of origin, treated as a random effect to account for variation in gene expression between patients. The Satterthwaite approximation of degrees of freedom, implemented in the lmerTest R package, was used to determine statistical significance of the fixed effect term in each peak-gene model tested.[55,56,142] To correct for multiple testing in each condition, the Benjamini-Hochberg method was applied using the *p.adjust()* function from the stats R package and peak-gene pairs with FDR-adjusted p-value < 1e-04 were deemed statistically significant peak-to-gene associations (Figures 2D, S13A, and 5C; Table S4, S7, and S8).[142,151] These were the final peak-to-gene association analyses performed for the cell line cohort dataset. However, for the Basal-like and Luminal subtype cohorts, intronic or distal intergenic peak-gene pairs with a significant association in at least one condition ("cancer" or "normal") were carried forward to a second phase of peak-to-gene association analyses to identify changes in peak-to-gene effect size between conditions.

To quantify the change in peak-to-gene effect size between conditions in a differential peak-to-gene association analysis, we combined the patient-specific metacells from both conditions and fit another LMM with an interaction term.[55,56] More specifically, gene expression was modeled as a function of peak accessibility, condition, the interaction between peak accessibility and condition, and patient of origin. All terms were treated as fixed effects, except for patient of origin which was treated again as a random effect to account for variation in gene expression between patients. As performed in the first phase of peak-to-gene association analyses, the Satterthwaite approximation of degrees of freedom was used to determine statistical significance of the fixed effect terms in each peak-gene model tested and the resulting p-values for the interaction term were corrected for multiple testing using the Benjamini-Hochberg method implemented in the *p.adjust()* function from the stats R package.[55,56,142,151] Peak-to-gene associations with FDR-adjusted p-value < 1e-04 for the interaction term were deemed statistically significant differential peak-to-gene associations (Figures 2E and 2F; Figure S13B-C; Tables S4 and S7).

Within the Basal-like and Luminal subtype analyses, significant differential peak-to-gene associations were visualized in a scatter plot of effect sizes for each condition using ggplot2 (Figures 2E, 2F, S13B, S13C, S6, and S14).[145] The cancer-specific peak-to-gene associations involving genes upregulated in the cancer condition were visualized in a heatmap of effect sizes for each condition using the *Heatmap()* function from ComplexHeatmap (Figures 3A and 4C).[141,150] The same visualization was performed for the putative silencer-to-enhancer switching events as well as normal-specific peak-to-gene associations involving genes upregulated in the normal condition (Figures S7A, S7C, S7F, S15A, S15C, and S15F). Select cancer-specific peak-to-gene associations of interest were visualized in a genomic browser track format, using ArchR's *plotBrowserTrack()* function, to display the ATAC-seq coverage patterns of the surrounding locus stratified by condition (Figures 3C, S11, 4E, and S18).[50] The same cancer-specific peak-to-gene associations of interest were visualized in scatter plots of peak accessibility by the inferred level of gene expression in metacells using ggplot2 (Figures 3D and 4F).[145]

Overlap analyses of genomic coordinates

To identify peak-to-gene associations that overlapped with existing ENCODE annotations downloaded from https://screen.encodeproject.org, the genomic coordinates of the peaks participating in the peak-to-gene associations were converted into a *GRanges* object using the GenomicRanges R package.[71,72,142,152] The genomic coordinates of the existing ENCODE annotations were also converted into a second *GRanges* object.[152] These two *GRanges* objects were used as input into the *subsetByOverlaps()* function from the IRanges R package.[142,152] The output from this function was a *GRanges* object containing the genomic coordinates of peaks that overlapped with the genomic coordinates of existing ENCODE annotations and was used to annotate the initial set of peak-to-gene associations for overlap with existing ENCODE annotations (Figures 2D, S13A, and 5C).

To perform the overlap analyses of putative enhancers between *in vitro* and *in vivo* BC cells for each subtype, the genomic coordinates of the putative enhancers *in vitro* and *in vivo* were converted into *GRanges* objects.[152] These two *GRanges* objects were used as input into IRanges' *findOverlaps()* function to identify the number of overlapping, or shared, putative enhancers between *in vitro*

and *in vivo* BC cells.[152] The same *GRanges* objects were used as input into IRanges' *subsetByOverlaps()* function, with the invert parameter set to TRUE, to identify the numbers of putative enhancers specific to *in vitro* and *in vivo* BC cells (Figures 5D–5O).[152]

To perform the overlap analyses of putative enhancer-target gene pairs between *in vitro* and *in vivo* BC cells for each subtype, the genomic coordinates of the putative enhancers participating in the putative enhancer-target gene pairs *in vitro* and *in vivo* were converted into *GRanges* objects.[152] These two *GRanges* objects were used as input into IRanges' *findOverlaps()* function to return a *GRanges* object containing the indices of genomic coordinates that overlapped between the two input *GRanges* objects.[152] The indices of overlapping genomic coordinates from each *GRanges* object were used to merge both sets of putative enhancer-target gene pairs, *in vitro* and *in vivo*, into one matrix based on overlapping putative enhancers. These sets of putative enhancer-target gene pairs, *in vitro* and *in vivo*, were screened for those with overlapping putative enhancers that linked to the same gene in both *in vitro* and *in vivo* settings. The resulting set of shared putative enhancer-target gene pairs in both settings was used to annotate the overlap status for the initial sets of putative enhancer-target gene pairs identified in each setting (Figures 5D–5O).

Gene set enrichment analysis

To perform gene set enrichment analysis with Hallmark gene sets from MSigDB, we inputted the list of genes of interest into the *enricher()* function from the clusterProfiler R package to test for significant enrichments via hypergeometric tests.[76–78,142] Gene sets with Benjamini-Hochberg adjusted p-values < 0.05 were deemed statistically significant enrichments.[151] The top three most significantly enriched gene sets were visualized using the *dotplot()* function from the enrichplot R package with the showCategory parameter set to "3" (Figures 3B, 4D, 5F, 5G, S7B, S7D, S7G, S15B, S15D, and S15G).[153]

Transcription factor (TF) motif analysis

To perform TF motif analysis for each differential association class (comprised of putative enhancer and/or silencer regions), we inputted the genomic coordindates of each set of peaks into the *find_motifs_genome()* function from the marge R package.[142,154] Note that the number of unique peaks in each differential association class were randomly downsampled to match the number of unique peaks in the smallest differential association class before computing motif enrichments. This was performed to avoid technical differences in motif enrichment between differential association classes due to large differences in initial sample size of peaks. The top 10% most variable motif enrichments (–log10(p-value)) across eight differential association classes were visualized in a heatmap using the *Heatmap()* function from the ComplexHeatmap R package (Figures S8A and S16A; Table S5).[141,142,150] The same procedure was used for the comparison of motif enrichments between putative cancer-specific and normal-specific enhancers (Figures S8B and S16B; Table S5).

Survival analysis

CALGB 40603 clinical data were acquired from dbGaP (phs001863.v1.p1) and upper quartile normalized RNA-seq expression data were downloaded from GEO (GSE154524). The normalized expression values were further log2-transformed.

FUSCC clinical data were downloaded from Table S1 of Jiang et al 2019 and RNA-seq fastq files were acquired from the NCBI Sequence Read Archive (SRP157974).[110] RNA-seq fastqs were aligned to the hg38 reference genome with STAR 2.7.6a and quantified with Salmon 1.4.0. Salmon counts were then upper quartile normalized and log2(x+1) transformed.

METABRIC clinical data and normalized expression data were acquired from the European Genome-Phenome Archive (EGAS00000000083) and came from Curtis et al 2012.[111] The normalized expression values were further log2(x+1) transformed. PR IHC data and HER2 FISH data were missing from this dataset, so TNBC samples were defined as IHC ER negative, no HER2 SNP6 gain, and HER2 SNP6 loss or HER2 negative by expression. Correspondingly, HR+/HER2- samples were defined as IHC ER positive, no HER2 SNP6 gain, and HER2 SNP6 loss or HER2 negative by expression. Stage 0, stage 4, and untreated samples were excluded from analysis.

SCAN-B clinical data and gene-level FPKM RNA-seq expression data were downloaded from Mendeley Data (https://data.mendeley.com/datasets/yzxtxn4nmd/3) and came from Staaf et al 2022.[112] The FPKM expression values were further upper quartile normalized and log2(x+1) transformed. Untreated samples, bilateral samples, multi-centric samples, lymph node samples, and normal samples were excluded from the analysis. Furthermore, sample duplicates were excluded (keeping the specimen that had the most frequently used library protocol, sequencer serial, library barcode, or pool name in the dataset, respectively).

TCGA-BRCA clinical data and RNA-seq fastqs were acquired from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/projects/TCGA-BRCA).[113] RNA-seq fastqs were aligned to the hg38 reference genome with STAR 2.7.6a and quantified with Salmon 1.4.0. Salmon counts were then upper quartile normalized and log2(x+1) transformed. Only fresh frozen tumor samples were considered for analysis, and stage 4 samples were excluded.

Cox proportional hazards models were fit using the survival R package for each of the 829 unique genes participating in 7,167 cancer-specific peak-to-gene associations, identified in the Basal-like subtype analysis, over the set of TNBC patients in each dataset (Table S6). The same was performed for each of the 288 unique genes participating in 1,931 cancer-specific peak-to-gene associations, identified in the Luminal subtype analysis, over the set of HR+/HER2- samples in each dataset (Table S6). Separate models were fit using each definition of survival available for a dataset. The survival R package was used for the Cox proportional hazard modeling with the formula Surv(time, event) ~ gene_expression.

Kaplan-Meier plots were created using the survminer R package, using median gene expression values as the "high expression" vs. "low expression" cutoffs, with log-rank P-values displayed (Figures S10A and S17A).

CNV landscape plots

TCGA-BRCA GISTIC2.0 gene-level copy number data (all_data_by_genes.txt) were downloaded from TCGA Firebrowse (http://firebrowse.org/?cohort=BRCA#).[113] Gene-level copy number scores were then converted to 534 pre-determined chromosomal segment copy-number scores by calculating the mean score of genes overlapping a given segment. The 534 segments used included chromosomal regions associated with cancer and whole-arm chromosome segments, as described in detail in Xia et al 2019 .[114] Segment-level copy number scores > 0.3 in a sample were considered copy number gains, and segment-level copy number scores < -0.3 in a sample were considered copy number losses. The percentage of all TNBC TCGA samples with a copy number gain/loss call for each segment was calculated and plotted by the relative segment order, with *HEY1* labeled between the two non-whole-arm segments it was closest to (chr8:62174237-62716885.BeroukhimS5.amp and chr8:81242335-81979194.BeroukhimS2.8-q21.13.amp) (Figure S10B). Correspondingly, the percentage of all HR+/HER2- TCGA samples with a copy number gain/loss call for each segment was calculated and plotted by the relative segment order, with *CRABP2* labeled between the two non-whole-arm segments it was closest to (chr1:151026302-152973244.BeroukhimS5.amp and chr1:158317017-159953843) (Figure S17B).

Correlation between copy number and expression

TCGA-BRCA clinical, expression, and copy number data were acquired and processed in the same way as described in the survival analyses and CNV landscape plotting. To select genes surrounding *HEY1* and *CRABP2*, the UCSC Human Gene Sorter tool (https://genome.ucsc.edu/cgi-bin/hgNear) was used. The 10 closest upstream and 10 closest downstream genes of HEY1 and CRABP2 by genome position that also had quantified TCGA gene expression data and copy number data available were chosen to be plotted, ordered by relative proximity to *HEY1* and *CRABP2*, respectively. The log2 transformed, upper-quartile normalized expression of each gene across TCGA TNBC and HR+/HER2- samples are plotted in Figures S10C and S17C, respectively. For the boxplots, the center line represents the median value, box limits represent upper and lower quartiles, and whiskers represent the 1.5x interquartile range. Any points outside of these ranges represent outliers. The Wilcoxon rank sum test was used to compare the expression of each nearby gene to the expression of *HEY1* and *CRABP2*, respectively. The Pearson correlation between the log2 transformed, upper-quartile normalized expression and the GISTIC2.0 copy number score of each gene across TCGA TNBC and HR+/HER2- samples are plotted in Figures S10D and S17D, respectively. Note that for *MIR5708* and *RN7SL107P*, the Pearson correlation could not be calculated because the expression was 0 for all samples.

Prioritization scheme for selection of *HEY1* and *CRABP2*

To screen for clinically relevant genes regulated by putative cancer-specific enhancers, we carried out a 7-step procedure to filter candidates in both the Basal-like and Luminal subtype analyses (Figure S9). First, we identified statistically significant peak-to-gene associations in either cancer or normal conditions with FDR-adjusted p-value < 1e-04. Next, we identified statistically significant differential peak-to-gene associations with FDR-adjusted p-value < 1e-04. Thirdly, we screened statistically significant differential peak-to-gene associations for those that showed a significant effect size > 0 in the cancer condition. Next, we screened these 'cancer-specific' peak-to-gene associations for those that involved genes upregulated in the cancer condition relative to the normal condition (FDR-adjusted p-value < 0.05 & log2FC $\geq$ 0.58). To prioritize prognostic genes, we screened for cancer-specific peak-to-gene associations involving genes that showed a significant hazard ratio > 1 in a majority of the external patient datasets tested (Cox p-values < 0.01). Finally, we prioritized cancer-specific peak-to-gene associations involving genes affected by copy number amplification events in >50% of TCGA TNBC or HR+/HER2- patients. This prioritization scheme revealed *HEY1* in the Basal-like subtype analysis and a total of five genes in the Luminal subtype analysis. These five genes were ranked by level of expression in scRNA-seq, and *CRABP2* was selected as one of the most highly expressed genes.