Genome Medicine

RESEARCH

Open Access

# Characterizing heterogeneous *cis*-regulatory elements in gene regulatory programs associated with breast cancer

Chisa Hori[1,2†], Kohei Kumegawa[3*†], Sumito Saeki[1,4], Yoko Takahashi[4], Liying Yang[1], Tomoyoshi Nakadai[1], Kazutaka Otsuji[3], Chikako Takahata[4], Yukinori Ozaki[5], Natsue Uehiro[4], Yurie Haruyama[4], Tomo Osako[6], Toshimi Takano[5], Seiichi Mori[7], Tetsuo Noda[8], Satoshi Fujii[2], Shinji Ohno[9], Takayuki Ueno[4] and Reo Maruyama[1,3*]

## Abstract

**Background** *Cis*-regulatory elements (CREs) control oncogene expression and malignant phenotypes. The high clinicopathological heterogeneity of cancer cannot be explained by gene expression alone, being attributed to CRE heterogeneity. However, characterizing cancer-associated CREs is challenging. To address this issue, we performed a single-cell epigenomic analysis of clinical specimens.

**Methods** To map the multicellular ecosystem of breast cancer (BC) and identify candidate CREs (cCREs), we performed a single-cell assay for transposase-accessible chromatin with sequencing (scATAC-seq) of 38 prospectively collected BC samples with various clinicopathological characteristics.

**Results** First, we performed single-cell chromatin accessibility profiling of a high-quality set of 22,775 cells from BC samples. Cells were annotated using marker gene accessibility and integration with existing single-cell RNA sequencing data. The chromatin accessibility patterns exhibited by cancer cells were consistent with the clinicopathological features of each tumor. We identified 224,585 cCREs across the BC ecosystem. By identifying cluster-specific differentially accessible cCREs (DA-cCREs) and constructing a putative enhancer–promoter network, we mapped the *cis*-regulatory landscape for cancer cells and the tumor microenvironment. The accessibility of putative enhancers targeting the same gene differed within or between tumors, highlighting intra- and inter-*cis*-regulatory tumor heterogeneity.

**Conclusions** This study provides a valuable resource for future epigenetic research on BC and highlights the diverse regulatory landscape within or among tumor(s), suggesting that cCREs regulate intra- and intertumor heterogeneity.

**Keywords** *Cis*-regulatory element, Single-cell ATAC-seq, Breast cancer, Tumor heterogeneity

†Chisa Hori and Kohei Kumegawa contributed equally to this work.

*Correspondence:
Kohei Kumegawa
kohei.kumegawa@jfcr.or.jp
Reo Maruyama
reo.maruyama@jfcr.or.jp
Full list of author information is available at the end of the article

## Background

Consortium studies, such as ENCODE and Roadmap Epigenomics, have advanced the knowledge on functional genomic elements [1, 2]. These initiatives have shed light on *cis*-regulatory elements (CREs), noncoding DNA sequences that regulate gene expression in development, differentiation, and disease. CRE (promoters, enhancers, silencers, insulators, etc.) functions are mediated by transcription factors (TFs) that bind CREs in a particular cellular context [3]. Since CREs function in a context-dependent manner, research on their functions and significance is challenging; however, recent studies have characterized CREs in various developmental processes [4–7].

Particularly in cancer tissue, characterizing CREs is more challenging than in normal tissue or during development, as cancer involves various genomic and epigenetic abnormalities, including structural variations and dysregulated expression of TFs and epigenetic modifiers [8–11]. These abnormalities complicate and diversify the context governing cancer-associated CREs. In fact, the epigenetic landscape of human tumor tissue exhibits larger interindividual variability than that of normal tissue [12, 13]. In addition, tumors not only consist of cancer cells but also of cells from the tumor microenvironment (TME) [14]. This can lead to the analysis of average CRE activity signals from both cancer and noncancer cells, resulting in a vague interpretation of CRE characteristics in the cancer ecosystem. Considering these difficulties, although the use of cell lines or animal models for characterizing cancer-associated CREs has advantages for experimental intervention, the function of CREs in tumor tissue could be different from that of these models.

A single-cell assay for transposase-accessible chromatin with sequencing (scATAC-seq), which captures cell diversity, identifies rare or unknown cell types, and helps create comprehensive CRE catalogs, can help overcome this hurdle of bulk assays [15]. scATAC-seq predicts a relationship between CREs, enabling the inference of enhancer-promoter networks (EPNs) [16]. In cancer samples, scATAC-seq can classify cells as cancerous or TME cells and distinguish the epigenetic heterogeneity of the cancerous cells within a tumor [17, 18]. After cell clustering, we can detect CREs per cell cluster, including cancer cells and various cell types from complex cell mixtures such as those obtained from clinical samples. Then, CREs can be cataloged, and the relationships between promoter CREs and other CREs can be evaluated to establish EPNs in cancer.

Breast cancer (BC) is a heterogeneous disease. Although clinical management strategies for BC are determined based on the expression states of hormone receptors, including the estrogen (ER) and progesterone receptors (PGR), and human epidermal growth factor receptor 2 (HER2), therapeutic outcomes sometimes vary among patients with the same receptor status [19–24]. A possible source of this heterogeneity may lie in divergent activity patterns of CREs among patients. Although recent studies have made significant progress in uncovering CRE heterogeneity in breast cancer [25–28], much remains to be understood.

Here, we performed scATAC-seq on 22 new clinical samples, in addition to 16 previously reported cases [29]. In our previous study, we focused on intratumor epigenetic heterogeneity, emphasizing the different patterns of key TF-binding motif enrichment within a single tumor. Specifically, an ER+/HER2− tumor contains two cancer cell populations: one with increased accessibility of ER binding sites and the other with decreased accessibility of ER binding sites and increased accessibility of GRHL2 binding motifs; however, we did not focus on individual CREs and the heterogeneity of CREs. In contrast, this study focuses on CRE heterogeneity within and across tumors, connecting CRE activity patterns to the clinicopathological features of each sample. We generated a comprehensive CRE catalog comprising ~225,000 candidate CREs (cCREs) from 22,775 cells in 38 samples and constructed a putative enhancer-promoter network (pEPN) identifying sets of cCREs that potentially regulate the expression of each gene. pEPN analysis revealed that different cCRE combinations regulated the same gene in different cancer clusters or tumors. Our results comprise a comprehensive catalog of BC-associated CREs and demonstrate the diversity of the *cis*-regulatory landscape in BC.

## Methods

### Cohort of this study and clinical specimens

A total of 38 breast cancer specimens were prospectively collected at the Cancer Institute Hospital, Japanese Foundation for Cancer Research. Among them, 16 samples were previously reported in our earlier study [29], while 22 samples were newly obtained for the present analysis. The specimens were collected from primary tumors, lymph node metastases, ascitic fluid, pleural effusion, and locally recurrent tumors. Primary, lymph node, and recurrent lesions were obtained via core needle biopsy of surgically resected tumors, whereas ascitic and pleural fluid samples were collected from aspirated effusions. Single-cell ATAC-seq was performed for all samples [30], and bulk DNA sequencing was additionally conducted for the P210LR and P190 specimens. The specimens were dissociated into single cells using a MACS Tumor Dissociation Kit and gentleMACS dissociation (Miltenyi Biotec), following the manufacturer's instructions. For pleural effusion samples, the drainage fluid was collected

after trocar placement, centrifuged, and purified using a Red Blood Cell Lysis Solution lysis buffer (Miltenyi Biotec).

#### scATAC-seq library preparation

The scATAC-seq libraries were prepared according to the droplet-based combinatorial indexing method [31], with some modifications. The indexed-Tn5 transposase was prepared as previously described [32]. After segmentation with the custom indexed-Tn5, the library was prepared using the SureCell ATAC-Seq Library Prep Kit (Bio-Rad) and a ddSEQ Single-Cell Isolator system (Bio-Rad). Q5 HiFi DNA Polymerase and USER II Enzyme (both from NEB) were used for a subset of samples in the Barcode Reaction mix and Enzyme Reaction mix. Then, the complete scATAC-seq libraries were loaded at 1.5 pM on a NextSeq 550 platform (Illumina), and sequencing was performed using the following read protocol: Read 1, 118 cycles; i7 index read, 8 cycles; and Read 2, 40 cycles.

#### Bulk DNA-seq

Genomic DNA was extracted from tumor samples using the AllPrep DNA/RNA Mini Kit (QIAGEN), following the manufacturer's instructions. The sequencing library was prepared using the Nextera DNA Flex Library Prep Kit (Illumina) and sequenced on a NovaSeq 6000 system (Illumina).

#### scATAC-seq: preprocessing, quality control, and clustering

The hg19 genome assembly was used for all scATAC-seq analyses. The initial processing of scATAC-seq data to generate barcoded and aligned read bam files was performed using the ATAC-Seq Analysis Toolkit (Bio-Rad). Sequencing data were then processed by converting BAM files into fragment files using Rsamtools (https://bioconductor.org/packages/Rsamtools).

ArchR [16] (v1.0.2) was used for further analysis unless otherwise noted. Fragment files were loaded using the createArrowFiles function. Quality control analysis was conducted for each cell; cells with a transcription start site (TSS) enrichment score < 4 for all samples were excluded. Further, cells were filtered based on the number of unique fragments, applying a sample-specific cutoff. This approach was necessary due to differences in sequencing depth across samples. Therefore, the cutoff for unique fragments was 1500–3000 per cell, depending on the sample. After creating arrow files for each sample, an ArchR project was generated containing all samples. To filter out doublets, addDoubletScores with $k = 10$, knnMethod = "UMAP," LSIMethod = 1 was used. In this study, samples with ≥ 50 cells were used for further analysis.

As different scATAC-seq methods were used for newly added or previously reported samples, we estimated the need for batch effect reduction between newly added (mainly using droplet-based single-cell combinatorial indexing ATAC-seq: DSCI) and previously reported samples (using droplet-based scATAC-seq: DSC). We compared the clustering results using or not using Harmony [33]. For default clustering, the addIterativeLSI function, with a genome-wide 500-bp tile matrix, was used to calculate the iterative LSI information. We clustered cells using addClusters (wrapper function of Seurat's FindClusters [34]) and the parameter maxClusters = 40. To run the uniform manifold approximation and projection (UMAP), we used the addUMAP function. For Harmony-corrected clustering, the addHarmony function (ArchR's wrapper function) was used with the method designation (DSC or DSCI) as the batch group to generate Harmony-corrected reduced dimensions. Subsequently, we used the same clustering function with Harmony-corrected reduced dimensions.

For clustering using only nucleosome-free fragments, ArchR's getFragmentsFromProject function was used to obtain all fragments per sample. Then, nucleosome-free fragments (fragment length < 150 bp) were subset. Clustering was performed using the same functions as above.

#### scATAC-seq: cell annotation

For cluster-based cell annotation, the gene accessibility of canonical cell type markers per cluster was visualized using the plotBrowserTrack function, and the gene activity score was visualized via the plotEmbedding function with imputation weights using the addImputeWeights function. To identify cluster-specific genes, the getMarkerFeatures function was used to calculate statistical values, and the getMarkers function was used to identify significantly active genes in each cluster (Wilcoxon $FDR < 0.01$ and $Log2FC \geq 1$). To perform gene enrichment analysis for the significantly active genes of each cluster, we used the compareCluster function of the clusterProfiler package (v3.18.1). Reference gene lists (CellMarker Augmented 2021 and PanglaoDB Augmented 2021) were downloaded from the Enrichr website [35].

To calculate the "epithelial score" and "fibroblast score," we used *EPCAM*, *CDH1*, and all keratins as epithelial genes, as well as *COL1A1*, *COL1A2*, *COL3A1*, *DCN*, *LUM*, *FBLN1*, *FN1*, *ACTA2*, *TAGLN*, *PDGFRB*, and *FAP* as fibroblast genes. We used the addModuleScore function to calculate the scores, and the mean values were defined as the epithelial and fibroblast scores for each cluster.

Publicly available single-cell RNA-seq (scRNA-seq) data of the BC samples [36] were retrieved using the Read10X — function of the Seurat package (v4.0.1) [34].

Then, the CreateSeuratObject function was used to generate the Seurat object. We also downloaded the processed Seurat objects from Figshare and obtained the correspondence between cell IDs and the cell type annotation defined in the original paper [36]. For visualization, the Seurat's NormalizeData, FindVariableFeatures, ScaleData, RunPCA, FindNeighbors (1–50 dimensions), and RunUMAP (1–50 dimensions) functions were used. To project cell type annotations from the scRNA-seq data onto our scATAC-seq data, we performed gene activity-based integration using the addGeneIntegrationMatrix function in ArchR. The function aligned the gene activity profiles derived from chromatin accessibility (GeneScoreMatrix) with the reference scRNA-seq dataset via Seurat's FindTransferAnchors function. The scRNA-seq object containing predefined cell type annotations served as a reference for label transfer. As a result, each scATAC-seq cell received predicted annotations.

#### scATAC-seq: inferring copy number variation (CNV)

The epiAneufinder [37] (v1.0.3) was used to infer CNVs from the scATAC-seq data. The original epiAneufinder article shows that the number of fragments per cell expected by epiAneufinder is ≥ 10,000; our dataset was smaller (5765 fragments per cell, median). Thus, we considered that using the epiAneufinder with the default settings for our data would produce false negatives, especially for the amplified regions. To address this potential issue, we investigated the optimal window size for P210LR, from which we obtained the actual genomic data. Based on the ATAC signal of P210LR, we ran the epiAneufinder with 12 window sizes ranging among 100 kb, 500 kb, and 1–10 Mb. For each epithelial cell, the percentage of windows overlapping with the region amplified by bulk DNA-seq was calculated from the windows considered amplified by epiAneufinder, and the average for each window was used as the accuracy index. In addition, for each non-epithelial cell, the percentage of windows considered unchanged (no deletions or amplifications) out of all windows was calculated and the average used as the normal window ratio. Finally, the optimal window size was determined to be 1 Mb, as it had a high accuracy index and normal window ratio.

We then ran the epiAneufinder on all sample scATAC fragments with a window size of 1 Mb. Due to the sparsity of scATAC-seq data, we assumed that regions with apparent signal dropout did not represent true deletions. Therefore, we focused our analysis on copy number amplification events. A window recurrently detected in > 25% of the TME cells in ≥ 2 samples was considered a false positive. The window considered to be amplified in ≥ 25% epithelial cells per sample, excluding the false-positive window, was then considered as the final "tumor amplification window." The copy number (CN) amplification score was calculated as follows: the overlap with all tumor amplification windows was detected for all windows evaluated by the epiAneufinder in each cell. Then, the CN amplification score was calculated as the percentage of overlaps actually estimated as amplified in each cell.

#### scATAC-seq: cCRE identification and profiling

To identify cCREs, we first generated pseudobulk replicates using the addGroupCoverages function. We then performed a MACS2 [38] (v2.2.7.1) peak call using the addReproduciblePeakSet function with the parameter method = "q" and cutOff = 0.1. For peak call, we used the 48 clusters identified in the two-round clustering. The addPeakMatrix function was used to add the reproducible peak set to the ArchR project. To annotate cCREs, the annotatePeak function of the ChIPseeker [39] package (v1.26.2) was used with tssRegion = c(−3000, 3000), TxDb = TxDb.Hsapiens.UCSC.hg19.knownGene, and annoDb = "org. Hs.eg.db." To identify sparse peaks, we obtained the peak matrix using the getMatrixFromProject function and calculated the number of inserts and that of cells for each peak. Peaks with < 10 inserts or < 10 cells were identified as sparse peaks and removed from further analysis. Peaks for pseudobulk samples were retrieved as previously described [40]. To calculate the significant overlap between each peak set, we utilized Bedtools' Fisher function.

#### scATAC-seq: DA-cCRE identification and pEPN construction

We used the ArchR's getMarkerFeatures function for statistical testing and the getMarkers function to identify DA-cCREs with $FDR < 0.25$ and $Log2FC > 0$. After obtaining DA-cCREs, we removed the sparse peaks. To construct the pEPN, we first calculated the co-accessibility using the addCoAccessibility function with the option maxDist = 500,000. The getCoAccessibility function was used to obtain co-accessible peak pairs with corCutOff = 0.4 and resolution = 500. Then, sparse peaks were filtered out among all co-accessible peaks. After uniquely identifying the cCREs present in the promoter based on the ChIPseeker annotation, we identified the co-accessible cCREs with that promoter and filtered out the correlated cCREs on other promoters. Correlated cCREs overlapping with the amplification windows defined by the epiAneufinder were filtered out as well. From the remaining promoter-cCRE pairs, promoters with > 50 correlated cCREs, which may be potentially affected by the DNA CN gain,

were filtered out. Finally, a set of promoters and curated cCREs (the pEPN) was constructed.

#### scATAC-seq: predicting super-enhancers (SE) from ATAC-seq signals

ROSE (rank ordering of super-enhancers) [41, 42] (v0.1) was used for SE identification. The ArchR's getFragmentsFromProject function was used to obtain the ATAC fragments per cluster. ROSE was then performed using "python ROSE_main.py -g hg19." To predict SE targets, overlapping genes with the ±500 kb of SE were identified.

#### Copy number analysis for bulk DNA-seq

For quality control and adapter sequence removal, we used TrimGalore (https://github.com/FelixKrueger/TrimGalore), a Perl wrapper of Cutadapt and FastQC. The AmpliconSuite-pipeline (v1.2.2; https://github.com/AmpliconSuite/AmpliconSuite-pipeline), an end-to-end wrapper of BWA MEM [43], SAMtools [44], CNVkit [45], AmpliconArchitect [46], and AmpliconClassifie r [47], was then used to identify focal amplification and the common source of amplicons.

#### Publicly available data analysis

Processed scMultiomics data by Terekhanova et al. [48, 49] were downloaded from the Human Tumor Atlas Network website (https://humantumoratlas.org) via the SYNAPSE system. The scATAC-seq and scRNA-seq data were processed and integrated using ArchR (v1.0.2) [16] and Seurat (v4.0.1) [34] in R. For scATAC-seq data, fragment files were imported and converted to Arrow files using the createArrowFiles function. Doublets were detected with addDoubletScores and filtered using filterDoublets. Cells with a TSS enrichment score > 4 and > 2500 fragments were retained for downstream analysis. For scRNA-seq data, expression matrices were loaded using Read10X and converted to Seurat objects with CreateSeuratObject. Cells were filtered to retain those with > 500 and < 6000 detected genes and < 10% mitochondrial content. Barcodes common to both scATAC and scRNA datasets were identified and used to filter the ArchR project and the Seurat object. The gene expression matrices were transferred from the Seurat object to the ArchR project using ArchR's addGeneExpressionMatrix function, with annotations from geneAnnoHg38. After that, ArchR was only used for analysis. Dimension reduction was performed using addIterativeLSI on both the TileMatrix (chromatin accessibility) and GeneExpressionMatrix, followed by integration via addCombinedDims. UMAP projection and unsupervised clustering (up to 40 clusters) were conducted using addUMAP and addClusters, respectively. Peak calling was performed per cluster using addGroupCoverages and addReproduciblePeakSet, followed by the creation of a peak-by-cell matrix with addPeakMatrix. Cluster-specific peaks were identified using getMarkerFeatures, followed by the extraction of marker peaks with getMarkers (cutoff: $FDR < 0.25$ and $Log2FC > 0$). Peak-to-gene regulatory links were inferred with addPeak2GeneLinks based on the correlation between peak accessibility and gene expression within ±500 kb. Correlated peak-gene pairs were extracted using getPeak2GeneLinks() with defined thresholds (corCutOff = 0.4, FDRCutOff = 0.01).

Processed scATAC-seq data by Regner et al. were downloaded from GEO (GSE243526) [50]. The data were processed using ArchR in the same manner as described above. The fragment files were converted to Arrow files using createArrowFiles and merged into an ArchR project via ArchRProject. Cells with TSS enrichment > 4 and > 2500 fragments were retained for downstream analysis. Dimension reduction and clustering were performed identically using addIterativeLSI, addUMAP, and addClusters (maxClusters = 40). Peaks were identified per cluster using addGroupCoverages and addReproduciblePeakSet with MACS2, followed by addPeakMatrix. Cluster-specific peaks were identified with getMarkerFeatures. Peaks satisfying $FDR < 0.25$ and $Log2FC > 0$ were extracted using getMarkers. The peak-gene linkages were used directly from Supplementary Tables 4 and 7 of their publication.

## Results

#### Single-cell chromatin accessibility atlas of BC tissues

We generated scATAC-seq data for 38 prospectively collected BC samples: 16 samples from our previous report [29] and 22 new samples (Additional file 1: Table S1). Samples were collected from various sites, including primary lesions, metastatic lymph nodes, pleural effusion,

(See figure on next page.)

**Fig. 1** Single-cell map of chromatin accessibility in breast cancer samples. **A** Clinicopathological features of the 38 samples analyzed. Samples of local recurrence were obtained from the chest wall. The criteria for estrogen (ER) and progesterone receptors (PGR) positivity were an Allred score ≥ 3, while human epidermal growth factor receptor 2 (HER2) positivity was defined as a score of +3. **B**, **C**, **D** Uniform Manifold Approximation and Projection (UMAP) of 22,775 cells colored by clusters (**B**), samples (**C**), and cell types (**D**). **E** Genome track plot showing the aggregate chromatin accessibility profile at marker gene loci for each cluster. The bar plot on the left shows the proportion of samples in each cluster; the sample colors correspond to **A**

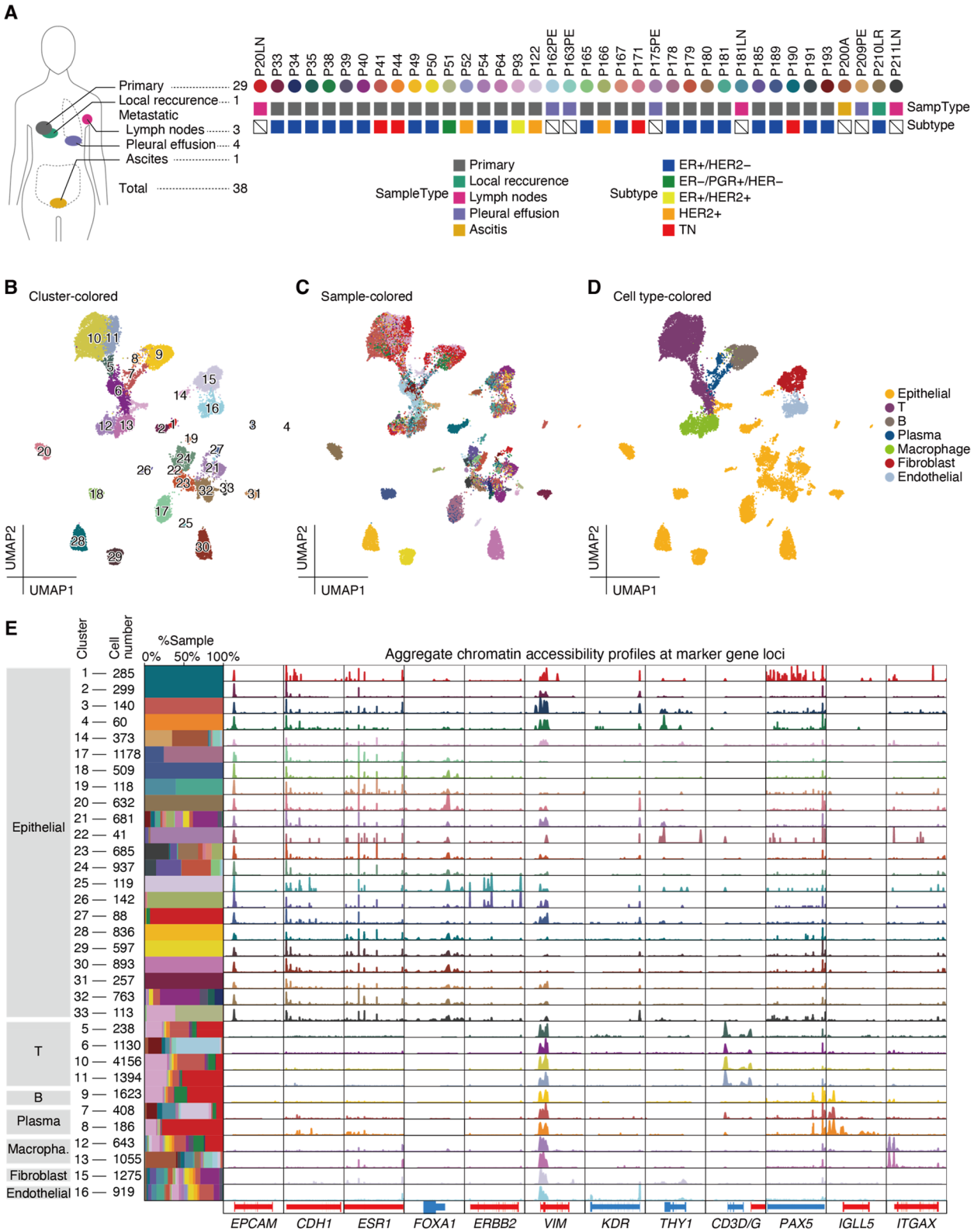


**Fig. 1** (See legend on previous page.)

ascites, and local recurrent lesions (Fig. 1A and Additional file 1: Table S1). After applying quality controls, single-cell chromatin accessibility profiles were acquired for 22,775 cells ("Methods") with an adequate fragment length density and TSS enrichment profiles (Additional file 1: Table S2; Additional file 2: Fig. S1A and B).

We next conducted iterative latent semantic clustering via ArchR [16], which yielded 33 clusters (Fig. 1B, C, D; Additional file 1: Table S3). The observed clustering was not attributable to scATAC-seq data quality (Additional file 2: Fig. S1C). Our previous dataset and the newly added one used slightly different scATAC methods; we added a sample index to the new samples ("Methods"). ArchR clustering can reduce batch effects by default; however, we further compared the results obtained with/without Harmony, a well-known batch effect control tool. Since we observed only minor differences in the clustering results with and without Harmony integration, we decided to use the results obtained without Harmony for subsequent analyses (Additional file 2: Fig. S1D, E, F, G, H). Further, there was variability in the fragment size distribution among samples (Additional file 2: Fig. S1A, B), potentially affecting clustering. To examine these potential effects, we performed clustering using only nucleosome-free fragments and confirmed minimal differences compared with those obtained using all fragments (Additional file 2: Fig. S1I, J). Therefore, we decided to adopt the ArchR clustering results for our downstream analysis.

To annotate the cell types in each cluster, we first visualized the chromatin accessibility of the coding regions of canonical cell type marker genes (Fig. 1E; Additional file 2: Fig. S2), identifying epithelial and endothelial cells, fibroblasts, T and B cells, plasmatic cells, and macrophages. A gene enrichment analysis of genes with significantly high activity scores for each cluster ($FDR < 0.01$ and $\log2FC \geq 1$; Additional file 1: Table S4) yielded results consistent with those obtained through marker gene-based annotation (Additional file 2: Fig. S3). Furthermore, we integrated publicly available single-cell RNA sequencing data from BC samples [36]. Based on the cell type annotation of the scRNA-seq data, we obtained a projected annotation for our scATAC-seq dataset, largely consistent with our gene activity-based annotation (Additional file 2: Fig. S4A, B, C, D). C14, derived from a metastatic ascites sample (P200A) and a metastatic pleural effusion sample (P209PE), exhibited reduced accessibility of epithelial markers (Fig. 1E), high activity genes marking epithelial and stromal cells, and a mixed scRNA-seq-based annotation of epithelial and fibroblast (Additional file 2: Fig. S4D). To ascertain the cell types in C14, we calculated two scores: the "epithelial score," which is defined by the module scoring of epithelial markers, and the "fibroblast score," which is defined by fibroblast markers ("Methods"). We classified C14 as epithelial cells because it exhibited an epithelial score comparable to other epithelial clusters while showing a distinctly lower fibroblast score compared to fibroblasts (Additional file 2: Fig. S5). This approach enabled us to obtain a comprehensive cell annotation for our scATAC-seq data (Fig. 1D; Additional file 1: Table S5). TME cell clusters contained cells from multiple samples, whereas epithelial cell clusters contained cells from a single or few samples (Fig. 1E), consistent with previous studies showing that TME cells are not as heterogeneous as epithelial/cancer cells [51–53].

Our dataset comprised various samples as sampling locations and receptor status (Fig. 1A; Additional file 2: Fig. S6A). The proportion of TME cells in each sample varied (Additional file 2: Fig. S6B). When evaluated by the collection site, the primary sample exhibited a higher percentage of epithelial cells than lymph node and pleural fluid ascites samples; additionally, the pleural fluid sample had a significantly higher percentage of macrophages than the primary sample (Wilcoxon $P = 0.041$; Additional file 2: Fig. S6C). Among primary tumors, comparisons among receptor subtypes revealed a potential trend for TN samples to contain a greater proportion of T cells than other subtypes (Additional file 2: Fig. S6D).

### The chromatin accessibility of cancer cells is associated with the clinicopathological features of the tumor of origin

We next subclustered the 9746 epithelial cells, identifying 35 epithelial clusters (hereafter, Ep1–Ep35) (Fig. 2A, B; Additional file 1: Table S6). To investigate the genomic abnormalities of the epithelial clusters, we inferred CNV from the chromatin accessibility using the epiAneufinder [37]. Given the relatively limited number of fragments in our dataset compared to the sequencing depth assumed by the epiAneufinder, we conducted bulk DNA sequencing on a single sample (P210LR) and utilized the resulting

(See figure on next page.)

**Fig. 2** Declustering of epithelial cells. **A**, **B** Uniform Manifold Approximation and Projection (UMAP) representing declustering of 9746 epithelial cells colored by epithelial clusters (**A**) and samples (**B**). **C** Dot plot showing bin-level log2 coverage by bulk DNA-seq for P210LR. Each dot represents a genome bin. **D** Heatmap showing scATAC-seq-based epiAneufinder's estimation of copy number variation (CNV). The genome window size is 1 Mb; 1508 windows are shown. **E** UMAP overlay of the CN amplification score. **F** Scatter plot showing the percentage of predominant samples and mean CN amplification scores for each epithelial cluster. **G** Heatmap showing the mean gene activity scores of breast cancer-associated genes. Table showing cluster annotations. **H** Volcano plot showing differentially active genes between breast cancer subtypes

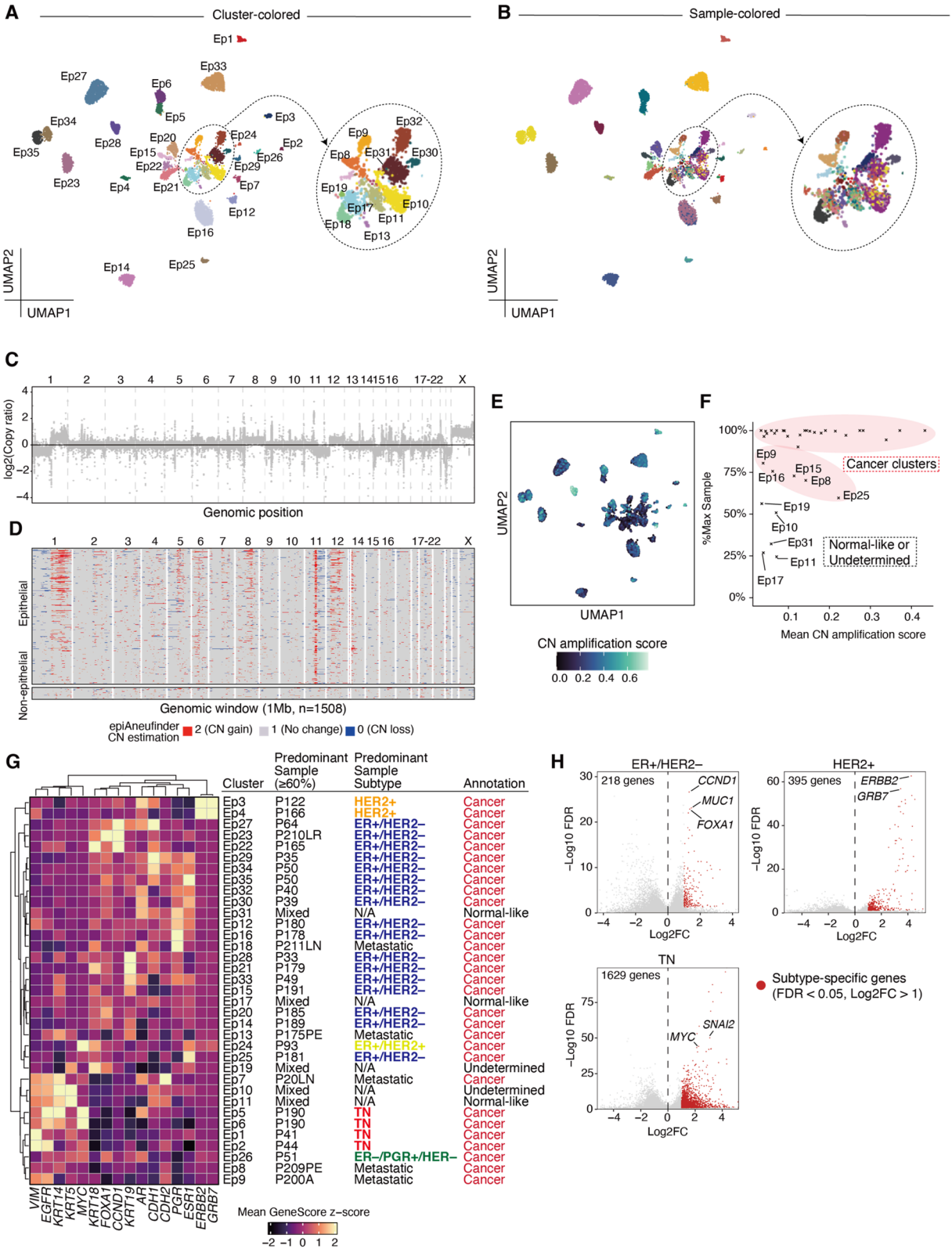


**Fig. 2** (See legend on previous page.)

data as ground truth to determine the optimal genomic window size to estimate (Additional file 2: Fig. S7A, B, C, D, E, F, G, H; "Methods"), identifying that the inferred CNV profile of the 1-Mb window was most consistent with the bulk DNA-seq result (Fig. 2C, D). To further validate the use of a 1-Mb window size, we compared the results of bulk DNA-seq and epiAneufinder analyses for sample P190. Bulk DNA-seq identified amplification of the MYC region in P190 (Additional file 2: Fig. S7I). In the scATAC-seq data, P190 was separated into two epithelial clusters, Ep5 and Ep6, which showed markedly different MYC gene scores (Additional file 2: Fig. S7J), suggesting the presence of distinct copy number clones within the tumor. When we applied epiAneufinder to the P190 sample using various window sizes, amplification of the MYC region in Ep5 and Ep6 was most clearly resolved with a 1-Mb window (Additional file 2: Fig. S7K). Based on these results, we performed CNV prediction using a 1-Mb window size. This analysis revealed recurrent gains in chromosomal arms 1q and 8q (Additional file 1: Table S7; Additional file 2: Fig. S8A, B). These alterations exhibited subtype-specific patterns: 1q gains were more prevalent in ER+/HER2− tumors, while 8q gains were more frequent in TNBC samples (Additional file 1: Table S7; Additional file 2: Fig. S8C, D).

Previous single-cell studies have shown that cancer cells frequently harbor copy number abnormalities and tend to form a single cluster within each tumor [51–53]. To characterize the epithelial clusters, we calculated the "CN amplification score" for all epithelial clusters (Fig. 2E; "Methods") and the proportion of the predominant samples for each cluster (Additional file 2: Fig. S8E) [27, 35]. Clusters consisting predominantly of cells derived from a single sample (> 90%) were identified and classified as cancer clusters, regardless of their CN amplification scores ($n = 25$: Ep1–7, Ep12–14, Ep18, Ep20–24, Ep26–30, Ep32–35; Additional file 1: Table S8). Ep25 had the lowest proportion of cells originating from its most frequent sample; however, this cluster included a mixture of cells from both a primary tumor and a lymph node metastasis from the same patient (59.7% from P181 and 40.3% from P181LN; Additional file 1: Table S8; Additional file 2: Fig. S8E), suggesting that Ep25 is a cancer cluster with shared chromatin accessibility between the primary and metastatic sites. Ep8 and Ep9 were composed of epithelial cells isolated from pleural effusion or ascites (Additional file 2: Fig. S8E). These clusters mainly contained cells from P209PE and P200A, along with smaller numbers of cells from P162PE and P163PE. These results imply that cancer cells in malignant pleural effusions and ascites share similar epigenomic profiles. Although the reason for the presence of cells from other samples is unclear, Ep15 and Ep16 were classified as cancer clusters because more than 70% of the cells originated from a single sample. We observed clusters with low CN amplification scores that included cells from multiple samples (Ep11, Ep17, and Ep31), which were possible normal cell clusters; however, these clusters were located near cancer clusters and did not show consistent copy number score patterns (Additional file 2: Fig. S8F). Thus, as these cells cannot be definitively classified as nonmalignant epithelial cells, we designated them as normal epithelial-like clusters. Clusters with low CN amplification scores but moderate levels of sample mixing were designated as "undetermined" (Ep10, 19). Therefore, we identified 30 cancer clusters, 3 normal epithelial-like clusters, and 2 undetermined clusters (Fig. 2F).

We next analyzed the gene activity scores within the epithelial clusters. Clusters originating from TNBC samples (Ep1, 2, 5, 6) showed elevated activity scores for basal lineage markers, such as *EGFR*, *VIM*, *KRT5*, and *KRT17* (Fig. 2G). Clusters from HER2+ samples (Ep3, 4) demonstrated high activity scores for *ERBB2* and *GRB7*, which are frequently coamplified. Most clusters from ER+/HER2− samples showed elevated activity scores for luminal lineage markers, such as *ESR1*, *FOXA1*, and *PGR*. Differential analysis of the gene activity score between BC subtypes showed that subtype-related genes were significantly activated (Fig. 2H; Additional file 1: Table S9). Taken together, these results indicate that our chromatin accessibility map is of high quality and accurately captures the characteristics of BC clusters.

### Identifying cCREs among BC ecosystems

To characterize the *cis*-regulatory profiles of our BC cohort, we used a cluster-specific and pseudobulk replicate peak calling method [16]. We called peaks from pseudobulk replicates from each cluster, which were not biased in the genomic annotation (Fig. 3A). As expected, the peaks of clusters belonging to the same subtype or cell type exhibited a high degree of overlap (Additional file 2: Fig. S9A). After merging the peaks and removing sparse peaks, we identified 224,585 reproducible, high-confidence cCREs (Fig. 3B; Additional file 1: Table S10; Additional file 2: Fig. S9B), the majority of which were non-promoter elements (70.6%). We then compared the cluster-based identified cCREs and the set of sample-based identified peaks to previously reported TCGA-BRCA ATAC-seq peaks [12] and found significant overlap (Additional file 2: Fig. S9C; "Methods"), highlighting the reliability of our dataset and the cCRE catalog.

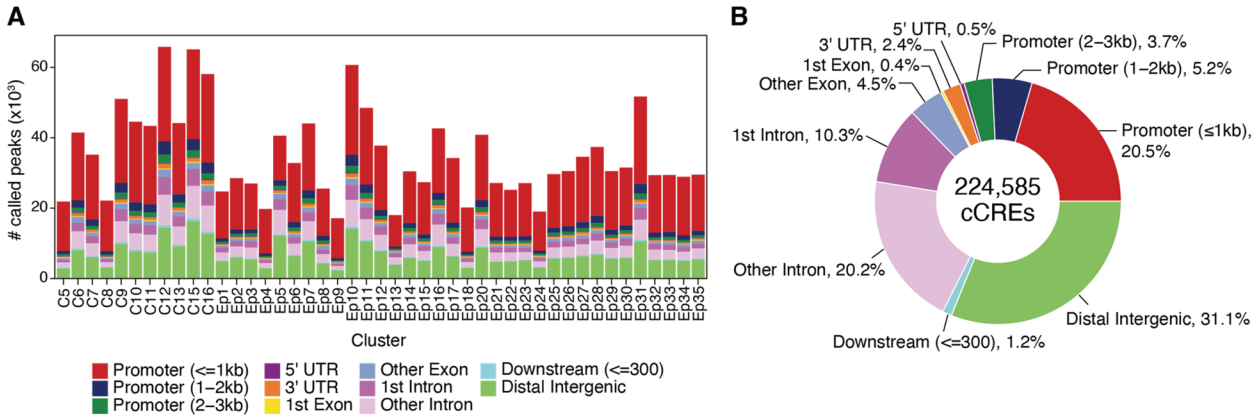


**Fig. 3** Candidate *cis*-regulatory elements (cCREs) identification. **A** Bar plot showing the number and genomic position of the called peaks for each cluster. **B** Genomic features of 224,585 reproducible, high-confidence peaks (cCREs)

## pEPN highlights the cis-regulatory landscape in the BC ecosystem

To explore the *cis*-regulatory landscape in the BC ecosystem, we identified differentially accessible cCREs (DA-cCREs) for each cluster (*FDR* < 25%; Additional file 1: Table S11; "Methods"). There was minimal overlap between DA-cCRE and the copy number gain regions estimated by epiAneufinder [37] in the predominant samples of each cluster (Fig. 4A; Additional file 2: Fig. S10A), suggesting a lower impact of genomic alterations on DA-cCREs. Motif enrichment analysis for DA-cCREs revealed that cell-type-specific TF motifs were enriched in DA-cCREs for some TME clusters: ERG and ELK4 in T-cell DA-cCREs (C10, 11), IRF8 in B/plasma cell DA-cCREs (C8, 9), and SPI1 in macrophage DA-cCREs (C12, 13) (Additional file 2: Fig. S10B). Although the enrichment of ER-responsive elements (EREs) was not the most prominent feature in ER+/HER2− clusters, we observed significant overlaps between the DA-cCREs of these clusters and ER ChIP-seq data from primary tumors [54] (Additional file 2: Fig. S10C).

To determine which genes are regulated by the DA-cCREs, we predicted the regulatory relationship between enhancers and promoters by building a pEPN (Additional file 2: Fig. S11A). We first conducted co-accessibility analysis across cCREs within a 500-kb distance (Fig. 4B), identifying 680,699 correlated cCRE pairs (correlation > 0.4; Fig. 4C). cCREs at the promoters of the same gene were made unique, resulting in a combination of 15,600 promoters and 260,826 cCREs. Of these cCREs, 124,235 cCREs located on promoters were filtered out. Since DNA copy number gain could be a strong driver of spurious peak-to-peak correlation, 34,208 cCREs overlapping with CN gain regions as defined by epiAneufinder were filtered out as well (Additional file 2: Fig. S11B). Indeed, promoters for the well-known amplified genes *ERBB2* and *MYC* had many correlated cCREs (Additional file 2: Fig. S11C). Of the remaining promoter-cCRE pairs, the promoter with the most correlated cCREs was CCND1;

(See figure on next page.)

**Fig. 4** Differentially accessible cCREs (DA-cCREs) and a putative enhancer-promoter network (pEPN) determine cluster characteristics. **A** Bar plot showing the number of DA-cCREs for each cluster. Red indicates the overlap with epiAneufinder estimated amplified windows. **B** Schematic of the computational strategy used to identify positively correlated cCRE pairs. **C** Density plot showing accessibility correlations between cCREs. Red indicates positively correlated cCRE pairs (correlation > 0.4). **D** Identification of correlated promoter-cCRE pairs and donut plot showing the promoter-correlated cCRE classification. The promoter in red indicates the cCREs on other promoters. CN gain (blue) indicates cCREs on epiAneufinder estimated amplified genomic windows. Extremely high number (green) indicates the cCREs whose correlated promoters have a large number (> 50) of the associated cCREs. pEPN (purple) indicates the remaining cCREs, called pEnhancers. **E** Density plot showing the distance distribution of the pEnhancers to each transcription start site (TSS). **F** Genome track plot showing *EPCAM* loci and its pEnhancers. The blue arch and red bar on the top of the track indicate the pEnhancers. Orange highlight indicates the *EPCAM* promoter. The bottom yellow track shows publicly available MCF7 H3K27Ac ChIP-seq (DRX013207) downloaded via ChIP-Atlas. **G**, **H** Dot plot showing the number of pEnhancers overlapping with the DA-cCREs of C11 (**G**) and C15 (**H**). Table showing the gene enrichment analysis for genes whose ≥ 5 pEnhancers overlapped with the DA-cCREs. **I** Heatmap showing the fraction of overlap between DA-cCREs for each cluster and pEnhancers for each target promoter. The five hierarchical clusters of the promoters specific to epithelial (1), T cell (2), B and plasma cell (3), macrophage (4), and endothelial and fibroblast (5) were further analyzed as shown in the right tables, which show the representative genes and gene enrichment analysis results obtained using PanglaoDB Augmented 2021

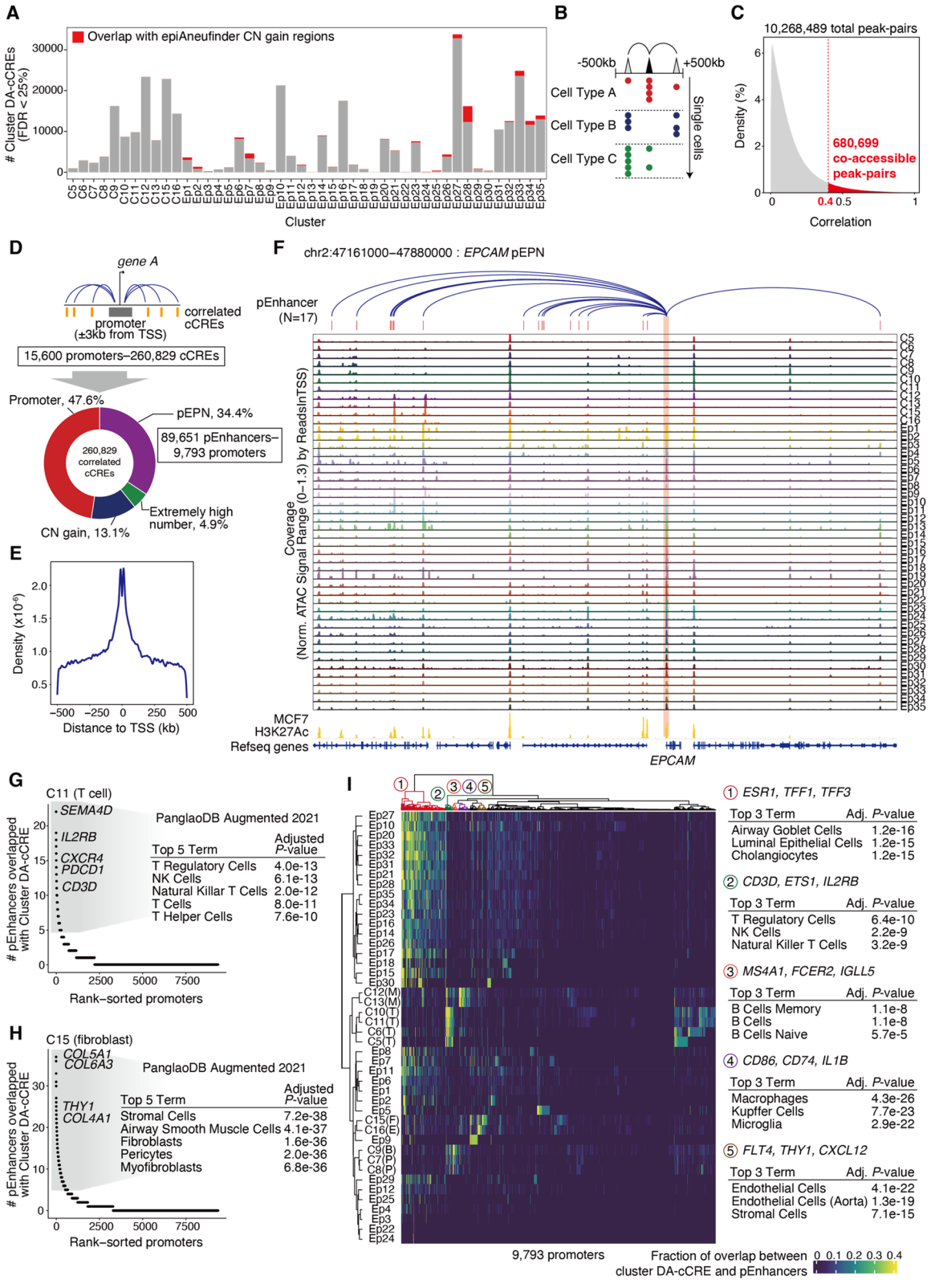


**Fig. 4** (See legend on previous page.)

genes in its vicinity also had many correlated cCREs (Additional file 2: Fig. S11D). The CCND1 locus was not considered a CN gain region by epiAneufinder, but adjacent genomic windows were (Additional file 2: Fig. S11E), suggesting that the epiAneufinder-based approach could overlook the amplified region. Thus, we filtered out 184 promoters with an extremely high number of correlated cCREs, potentially affected by DNA CN gain. Finally, we defined a pEPN including 9793 promoters and 89,651 putative enhancers (pEnhancers) (Fig. 4D; Additional file 1: Table S12). As expected, a density plot illustrating the distances between each peak and its target promoter showed a great decline with increasing distance between them (Fig. 4E). Most promoters were associated with < 10 distinct pEnhancers (Additional file 2: Fig. S11F), while most pEnhancers were predicted to interact with only a single promoter (Additional file 2: Fig. S11G). The promoter of *EPCAM*, an epithelial marker gene, had generally high accessibility in epithelial clusters; however, the accessibility of its pEnhancers ($N=17$) varied across clusters (Fig. 4F), implying *cis*-regulatory heterogeneity across tumors.

Using the pEPN, we determined the target genes for the DA-cCREs in each cluster. The number of overlaps between DA-cCREs and pEnhancers highlighted cell type-specific active pEnhancers and their target genes (Fig. 4G, H, I). ER+ epithelial clusters had a large overlap between DA-cCREs and pEnhancers targeting genes associated with luminal epithelial cells, including *ESR1*, *TFF1*, and *TFF3*; similarly, stromal and immune cell clusters had a large fraction of overlapping targeting genes associated with their cellular identity (Fig. 4I).

Finally, we compared pEPN and putative SEs detected by cluster-specific ATAC signals via ROSE [41, 42] (Additional file 2: Fig. S12A). The considerable number and length of the identified SEs led to the detection of a substantial number of genes (Additional file 2: Fig. S12B, C, D), rendering the method unsuitable for further cluster characterization. Taken together, these results suggest that the constructed pEPN is a reliable foundation for elucidating the *cis*-regulatory landscape of diverse cell types and cancer cell populations.

#### Comparison of our dataset with other scATAC-seq studies

To assess the robustness and generalizability of our analysis, we validated our results using independent single-cell chromatin accessibility datasets from previous studies: Terekhanova et al. [48] and Regner et al. [28]. After quality filtering, we obtained 87,390 high-quality scMultiome profiles from 14 breast cancer samples in the Terekhanova et al. dataset and 107,894 high-quality scATAC-seq profiles from 16 samples in the Regner et al. dataset (Additional file 2: Fig. S13A, B, C, D). We then performed peak calling, which yielded 572,181 peaks from the Terekhanova dataset and 525,922 peaks from the Regner dataset. Of the 224,585 cCREs identified in our dataset, 84.3% overlapped with peaks in the Terekhanova dataset and 78.3% with those in the Regner dataset (Additional file 1: Table S13; Additional file 2: Fig. S13E). In contrast, only 57.6% overlapped with bulk ATAC-seq peaks from TCGA-BRCA (Additional file 2: Fig. S9C), suggesting the superior sensitivity of single-cell data in capturing cell-type-specific regulatory elements.

We next evaluated the overlap of cluster-specific peaks across datasets. Immune cell-specific peaks (our clusters C5–C13) exhibited high Jaccard indices with clusters C28–C31 and C33–C34 from Terekhanova et al., as well as with clusters C1–C2, C19, and C21 from Regner et al. Fibroblast-specific peaks (our cluster C15) aligned with clusters C24–C25 (Terekhanova et al.) and C13–C18 (Regner et al.). Meanwhile, endothelial-specific peaks (our cluster C14) overlapped with cluster C23 (Terekhanova et al.) and clusters C10–C12 (Regner et al.) (Additional file 1: Table S14; Additional file 2: Fig. S13F). Epithelial cell clusters exhibited greater inter-patient heterogeneity, resulting in lower—though still detectable—overlap across datasets. These results suggest that the epigenetic landscapes of the breast cancer ecosystem are at least partially conserved across datasets.

To further validate our pEPN, we leveraged the single-cell multiomics structure of the Terekhanova dataset, which allows for peak-gene linkage analysis based on joint chromatin accessibility and gene expression profiles from the same cells. Peaks located within ±500 kb of each gene's TSS were retained if their accessibility showed a correlation greater than 0.4 with gene expression and an $FDR < 0.01$ ("Methods"). This yielded 276,400 putative enhancer-gene linkages involving 12,201 genes. Additionally, Regner et al. performed a similar analysis using a linear mixed-effects model and provided enhancer-gene linkages specific to basal-like and luminal cell states in their supplementary data. We compared these datasets with our pEPN, which contains 89,633 regions linked to 9793 genes. Of these, 2928 genes (29.9%) shared at least one putative enhancer with the Terekhanova dataset, 3847 genes (39.3%) shared at least one putative enhancer with the basal-like cell-specific links from Regner et al., and 3785 genes (38.7%) overlapped with at least one putative enhancer of the luminal cell-specific links (Additional file 1: Table S15; Additional file 2: Fig. S13G). These substantial overlaps suggest that core gene regulatory relationships are conserved across independent patient cohorts, thereby supporting the validity of our inferred pEPN.

Taken together, these results demonstrate the robustness and biological relevance of our dataset and analytical framework, underscoring their utility in characterizing

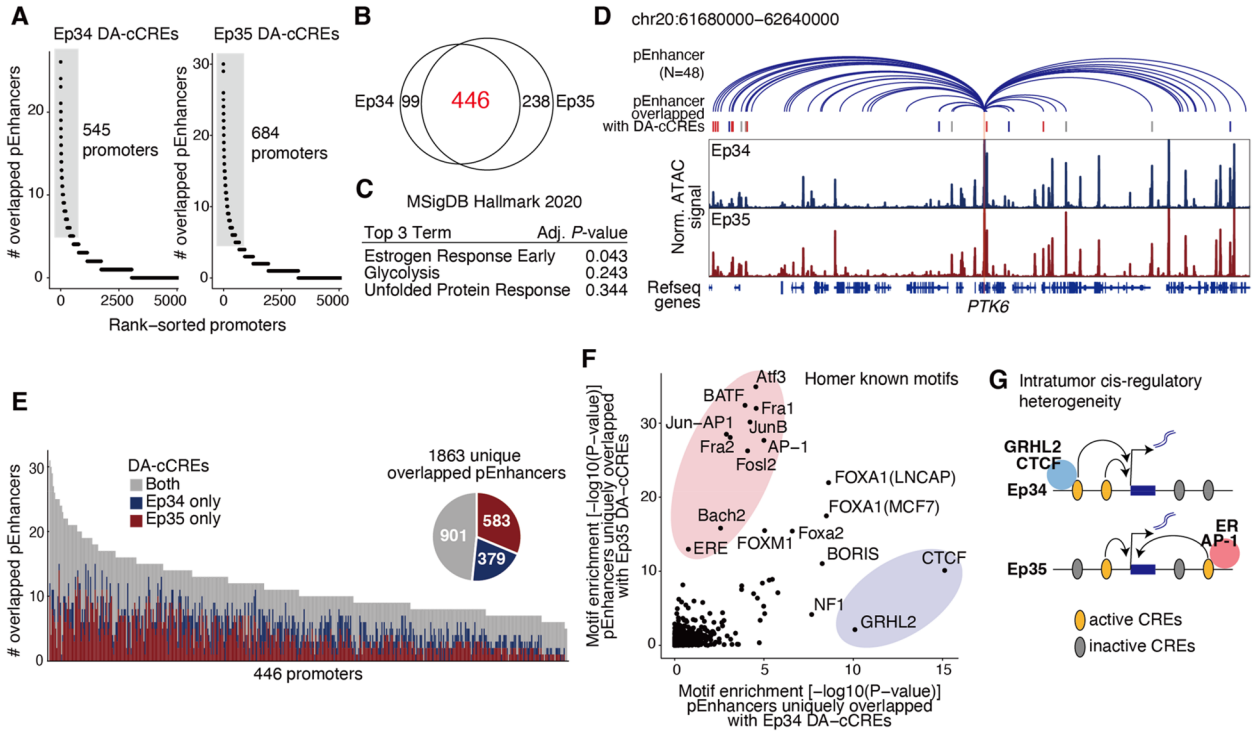


**Fig. 5** *Cis*-regulatory heterogeneity within a single tumor. **A** Dot plot showing the number of pEnhancers overlapping with the DA-cCREs of Ep34 (left) and Ep35 (right). The 545 and 684 promoters whose ≥ 5 pEnhancers overlapped with the DA-cCREs of Ep34 and Ep35. **B** Venn diagram showing the overlap between the 545 promoters of Ep34 and the 684 promoters of Ep35. **C** Gene enrichment analysis for the overlapping 446 genes. **D** Genome track plot showing the *PTK6* loci and its pEnhancers. The blue arch on the top of the track shows the pEnhancers. The bars represent DA-cCREs: gray indicates both Ep34 and Ep35 DA-cCRE, blue indicates Ep34 DA-cCRE, and red indicates Ep35 DA-cCREs. Orange highlight indicates the *PTK6* promoter. **E** Bar plot showing the number of pEnhancers overlapping with DA-cCREs. Gray indicates overlap with both DA-cCREs, blue with Ep34 DA-cCREs only, and red with Ep35 DA-cCREs only. Pie chart showing the number of unique pEnhancers overlapping with DA-cCREs. **F** Scatter plot showing the motif enrichment of the pEnhancers, which target the 446 promoters, overlapping with Ep34 or Ep35 DA-cCREs. **G** Schematics depicting the intratumor *cis*-regulatory heterogeneity

heterogeneous cis-regulatory programs in the breast cancer ecosystem.

### Intratumor cis-regulatory heterogeneity

We previously reported that P50, an ER+/HER2− tumor, has two distinct cancer clusters with different *cis*-regulatory properties, i.e., one having ERE enrichment and the other having GRHL2 motif enrichment [29]. Here, we aimed to characterize this intratumor *cis*-regulatory heterogeneity using pEPN. In this study, the cancer cells in P50 were divided into Ep34 and Ep35, exhibiting distinct patterns of motif enrichment (Additional file 2: Fig. S14). We identified 545 and 684 promoters with ≥ 5 pEnhancers overlapping with Ep34 DA-cCREs and Ep35 DA-cCREs (Fig. 5A). Of those, 446 of the promoters for Ep34 and Ep35 overlapped (Fig. 5B; Additional file 1: Table S16) and were significantly associated with estrogen response (Fig. 5C). *PTK6*, which is associated with poor BC outcome [55, 56], had 48 pEnhancers, of which six overlapped with both DA-cCREs, four only with Ep34 DA-cCREs, and nine only with Ep35 DA-cCREs (Fig. 5D). Of the 446 promoters, many pEnhancers overlapped both Ep34 and Ep35 DA-cCREs, while half of the pEnhancers overlapped only Ep34 or Ep35 DA-cCREs (Fig. 5E; Additional file 1: Table S17), suggesting that even for the same gene/promoter, some CREs are common to both clusters, while others are different. Motif analysis revealed that GRHL2 and CTCF motifs were enriched in pEnhancers overlapping with Ep34 DA-cCREs, and that ERE and AP-1 family motifs were enriched in pEnhancers overlapping with Ep35 DA-cCREs (Fig. 5F). Taken together, these results suggest the presence of intratumor *cis*-regulatory heterogeneity, driven by the activity of various TFs (Fig. 5G).

### Cis-regulatory heterogeneity across ER+/HER2− tumors

Our dataset included 17 ER+ cancer clusters (Fig. 2G); next, we focused on the *cis*-regulatory heterogeneity across ER+/HER2− tumors. We identified 538 promoters with ≥ 5 pEnhancers overlapping with the DA-cCREs

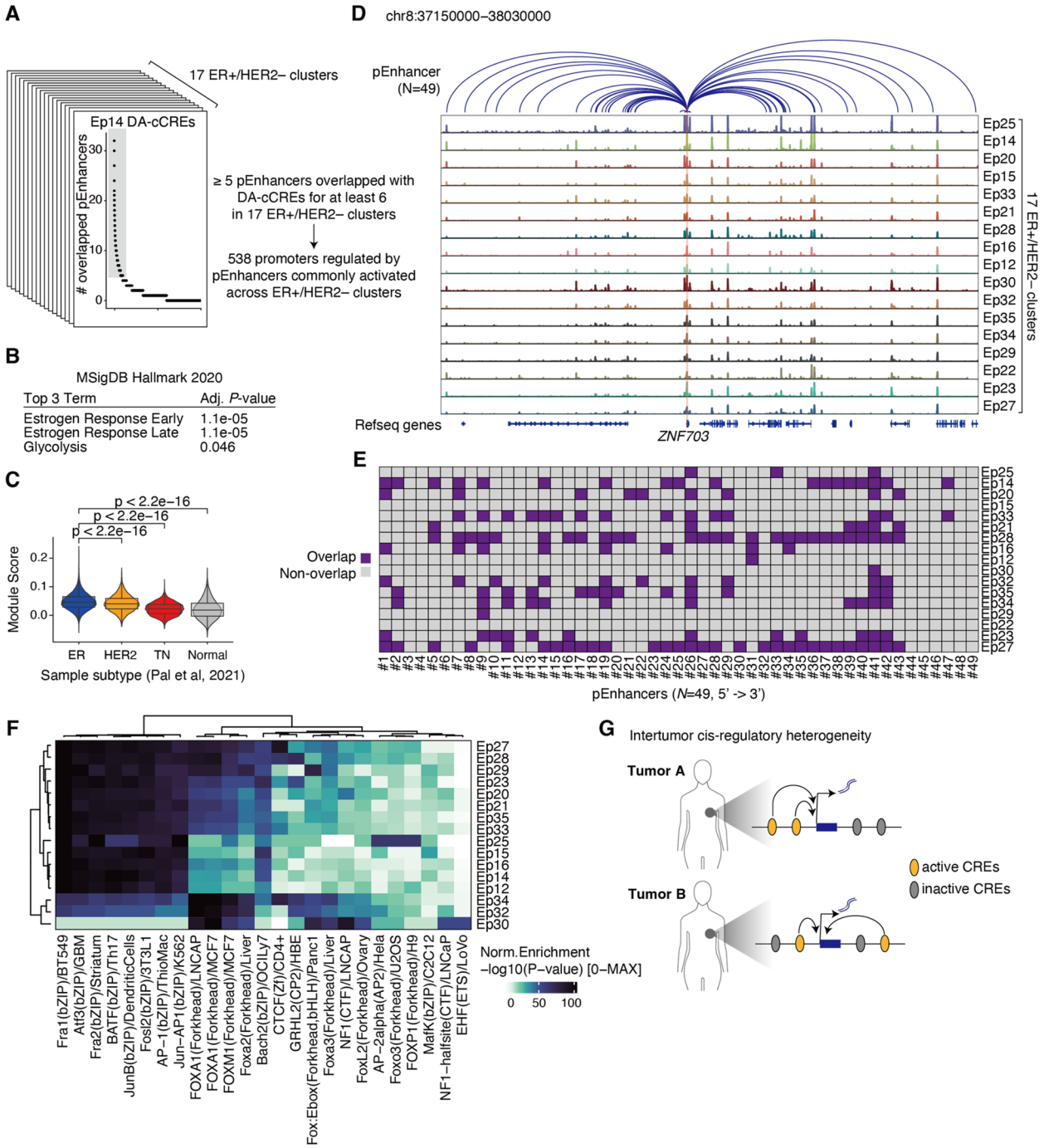


**Fig. 6** *Cis*-regulatory heterogeneity between ER+/HER2– tumors. **A** Schematic for the identification of promoters regulated by pEnhancers commonly activated across ER+/HER2– clusters (≥ 6 in 17 clusters). **B** Gene enrichment analysis of genes regulated by pEnhancers commonly activated across ER+/HER2– clusters (≥ 6 in 17 clusters). **C** Boxplot showing the module scores of commonly activated genes across BC subtypes in publicly available scRNA-seq data (Pal et al., 2021). *P*-values calculated by the Wilcoxon rank-sum test. **D** Genome track plot showing ZNF703, one of the commonly activated genes, and its pEnhancers. The blue arch on the top of the track shows the pEnhancers. **E** Heatmap showing the overlap between ZNF703 pEnhancers and DA-cCREs for each ER+/HER2– cluster. **F** Motif enrichment analysis of pEnhancers overlapping with the DA-cCREs for each ER+/HER2– cluster. The top 10 enriched Homer known motifs for each set of pEnhancers are shown. Ep22 is not shown because it has a single pEnhancer overlapping with its DA-cCREs. **G** Schematics depicting inter-tumor *cis*-regulatory heterogeneity

of ≥ 6 clusters among the 17 ER+ cancer clusters (> 33%) (Fig. 6A). As expected, these promoters/genes were significantly associated with the estrogen response (Fig. 6B). In the scRNA-seq dataset, these genes were significantly highly expressed in ER+ epithelial cells compared with other subtypes and normal epithelial cells (Fig. 6C), suggesting that these genes are commonly expressed in ER+/HER2− tumors. In addition to the intratumor *cis*-regulatory heterogeneity between Ep34 and Ep35, we observed diverse pEnhancers and DA-cCRE overlap patterns for each ER+/HER2− cluster (Fig. 6D, E; Additional file 1: Table S18). Motif analysis revealed that the AP-1 family and FOXA1 motifs were top-enriched in pEnhancers overlapping with DA-cCREs in each cluster; however, various motif enrichment patterns were observed (Fig. 6F). The results suggest that the *cis*-regulatory mechanisms governing the expression of genes commonly observed in ER+ tumors are diverse, and that this inter-tumor *cis*-regulatory heterogeneity may be due to subtle variations in TF activity (Fig. 6G).

## Discussion

The noncoding genome remains poorly characterized. However, numerous studies have demonstrated that CREs in the noncoding genome play an important role in gene expression and disease. Characterizing CREs in cancer is challenging because of their context-specific nature. Cataloging cancer-associated CREs and identifying key regulatory modules can provide insights into the mechanisms of cancer pathologies, such as carcinogenesis, metastasis, and recurrence, beyond gene expression patterns.

scATAC-seq can identify cCREs in complex tissues containing diverse cell types. However, determining their function is challenging. Although these are "candidate" CREs, indicating that a specific genome region is accessible in its context, it is unknown if they are functional. Determining the function of CREs identified by analyzing clinical samples in a specific model, such as a cell line or PDX, is insufficient for a fundamental understanding of CREs due to their context-dependent nature. In this study, we investigated the cellular properties of each cellular context (cluster) and constructed pEPNs, putative interactions between cCREs, to elucidate the *cis*-regulatory machinery in the different cell populations that comprise BC tissue. This strategy may exemplify how tumor-derived scATAC-seq data can be used to identify the CRE signatures and functions that define cellular properties in the cancer ecosystem.

By constructing pEPNs, we highlighted the diverse activity of CREs and suggested that different *cis*-regulatory mechanisms may apply to the same gene in different cell populations or tumors. In this study, the *cis*-regulatory heterogeneity in ER+ tumors was clear because of the high proportion of ER+/HER2− tumors in our cohort. While many patients with ER+ tumors respond to endocrine therapy and have a good prognosis, a few patients develop metastatic recurrence and are difficult to treat. Recently, this prognostic difference has been related to the epigenome. In our previous study, we found different chromatin accessibility patterns in ER+ BC despite homogeneous gene expression. In this study, we showed that the activity patterns of putative enhancers regulating a group of genes involved in ER signaling differ among cancer cells and between tumors. This *cis*-regulatory heterogeneity may affect ER+ BC evolvability and be the basis for future treatment of acquired resistance.

The present study is limited by the relatively small number of fragments and the small sample size. scATAC-seq data analysis is challenging due to its high sparsity. The limited number of fragments in this study may have resulted in the omission of crucial signals, as suggested by the inadequate genome amplification estimation by epiAneufinder. Moreover, this sparsity may have introduced vulnerability in our correlation-based inference of pEPN, potentially limiting the robustness of the predictions. Future studies would benefit from the development and application of more rigorous approaches for inferring gene regulatory networks, such as the advanced framework proposed by Regner et al. [28]. Our scATAC-seq analysis of 38 BC tissue samples, although including more samples than other studies [48, 57], was biased toward the ER+/HER2− subtype and did not provide a detailed examination of the *cis*-regulatory landscape of the TN and HER2+ subtypes. Furthermore, the specimens used were from core needle biopsies, with a relatively low number of cells per specimen. The limitations of this dataset preclude a deeper understanding of intratumor heterogeneity, including an investigation of differences in TME cell content. Furthermore, although we aimed to exclude the effect of genomic amplification, the relationship between genomic amplification and chromatin accessibility could affect tumor biology. The integration of genomic data and multimodal analysis could facilitate a more comprehensive understanding of the interrelations between the genome and epigenome.

Recently, scATAC-seq studies have been published analyzing a large number of clinical samples from various cancer types as part of the Human Tumor Atlas and the TCGA network [48, 57]. These cross-sectional analyses contributed significantly to our understanding of the regulatory mechanisms involved in cancer development and metastasis. However, it raises the following questions: How many samples are needed to reveal

the full extent of cancer-associated CREs? Our study identified approximately 225,000 cCREs, with several non-overlapping regions with TCGA-BRCA ATAC-seq peaks. A more complete map of the CRE genome requires the analysis of cases with different pathological features and complex clinical courses, including samples from patients undergoing different treatments and with different recurrence histories, to examine the association between the CRE landscape and clinicopathological features.

## Conclusions

Our study generated a comprehensive map of BC-associated CREs and linked them to the clinicopathological features of each tumor. Thus, our findings open new avenues for characterizing cancer-associated CREs, which have high context specificity and large heterogeneity among patients, providing a resource for BC epigenomic studies.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-025-01562-1.

Additional file 1: Supplementary Tables. Table S1. Clinicopathological features of 38 samples. Table S2. Quality metrics for each cell. Table S3. Cluster x Sample matrix (all clusters), related to Fig. 1. Table S4. Statistical analysis result for gene activity score (differentially active genes for each cluster). Table S5. Cell type annotation for each cell, related to Fig. 1B,D and Fig. S4I. Table S6. Cluster x Sample matrix (epithelial clusters), related to Fig. 2. Table S7. epiAneufinder-defined amplified genomic windows. Table S8. %Max sample and Mean CNV score for each cluster, related to Fig. 2F. Table S9. Statistical analysis result for gene activity score (differentially active genes for each subtype), related to Fig. 2H. Table S10. cCREs, related to Fig. 3. Table S11. DA-cCREs, related to Fig. 4A. Table S12. pEPN (9,793 promoters). Table S13. cCREs converted to hg38. Table S14. DA-cCREs converted to hg38. Table S15. pEnhancers converted to hg38. Table S16. Overlap promoters(genes) between Ep34 and Ep35 DA-cCREs overlapping with ≥5 pEnhancers (446 promoters), related to Fig. 6B. Table S17. # pEnhancers overlapped with Ep34 and/or Ep35 DA-cCREs, related to Fig. 6E. Table S18. Overlap status between pEnhancers of the 538 promoters and ER+/HER2- cluster DA-cCREs, related to Fig. 7 (overlap: 1, non-overlap: 0).

Additional file 2: Supplementary Figures. Fig. S1. Quality control of scATAC-seq data. (A) Fragment size distribution per sample. Sample colors correspond to Fig. 1A. (B) Normalized insertion profile around transcription start sites. (C) UMAP overlay showing TSS enrichment score, promoter ratio and log10(number of fragments). (D) UMAP colored by scATAC-seq method (DSC or DSCI), clustered without additional batch effect control (ArchR standard clustering). (E) UMAP stained by scATAC-seq method (left) and clusters, clustered with Harmony, a batch effect control tool. (F-G) The proportion of cells from the scATAC-seq method in each cluster without Harmony (F) and with Harmony (G). (H) The degree of mixing of DSC and DSCI methods in each cluster. P value was calculated by Wilcoxon rank sum test. (I) UMAP colored by clusters identified using only nucleosome-free fragments. (J) Heatmap showing the cell overlap rate between clustering using all fragments and using nucleosome-free fragments. Fig. S2. Gene activity of cell type markers. UMAP overlay showing gene activity scores of cell type marker genes. Fig. S3. Gene enrichment analysis of cluster-specific activated genes. (A–B) gene enrichment analysis using CellMarker Augmented 2021 (A) and PanglaoDB Augmented 2021 (B) for the reference. Fig. S4. Re-analysis of the scRNA-seq dataset (Pal et al., 2021) and projection of scRNA-seq onto our scATAC-seq data. (A) UMAP visualization of the scRNA-seq dataset comprising 27 samples, colored by the cell types from the original study. TAMs, tumor associated macrophages; CAFs, cancer-associated fibroblasts; DCs, dendritic cells. (B) Schematic illustration of the transferring cell type labels from the scRNA-seq data onto our scATAC-seq data. (C) UMAP of our scATAC-seq data colored by transferred cell type labels. (D) Proportions of transferred cell types in our scATAC-seq clusters. Fig. S5. Epithelial and fibroblast scores per cluster. (A) Dot plot showing the epithelial score (mean module scores of gene activity of *EPCAM*, *CDH1*, and keratins) per cluster. (B) Dot plot showing the fibroblast score (mean module scores of gene activity of the fibroblasts markers) per cluster. Fig. S6. Cell composition of the tumor microenvironment. (A) UMAP colored by sample type. (B) UMAP colored by sample type. (B) Bar plots showing the proportions of cell types per sample. (C–D) Box plots showing the proportions of cell type per sample type (C) or breast cancer subtype (primary sample only) (D). P-values were calculated by Wilcoxon rank sum test. Fig. S7. Optimization of the window size for epiAneufinder. (A–G) Heatmap showing the epiAneufinder copy number estimation for different genomic window sizes ranging from 100kb, 500kb, 2Mb, 4Mb, 6Mb, 8Mb, and 10Mb for P210LR sample. (H) Line graph showing the accuracy index and normal window rate for each window size. The black arrow indicates the most appropriate window size (high values of both of accuracy index and normal window ratio). (I) Dot plot showing bin-level log2 coverage by bulk DNA-seq for P190. Each dot represents a genome bin. The genomic bins containing MYC coding region were red-highlighted. (J) Boxplot showing the MYC gene activity score of Ep5 and Ep6 derived from P190. (K) Heatmap showing the epiAneufinder copy number estimation for different genomic window sizes ranging from 500kb, 1Mb, and 2Mb for P190 sample. Fig. S8. Inferred copy number amplification and sample mixing in clusters. (A) Karyotype plot showing the epiAneufinder-defined amplified regions in the cells of the tumor microenvironment. Red indicates the recurrently detected regions in at least two samples. (B–D) Karyotype plots showing the amplified regions defined by epiAneufinder in all epithelial cells (B), ER+/HER2– cancer cells with recurrent amplified regions highlighted in red (≥7 samples) (C), and TN cancer cells with recurrent amplified regions highlighted in red (≥2 samples) (D). (E) Bar plot showing the proportion of samples in each cluster. (F) Box plot showing the CN amplification score per cluster. Fig. S9. Peak profiling. (A) Heatmap showing the Jaccard similarity index between the peaks called in each cluster. The numbers to the right of the cluster name indicate the number of peaks. (B) Scatterplot showing the number of insertions (X-axis) and the number of cells (Y-axis) for each peak. Blue indicates a sparse peak. (C) Venn diagram showing the overlap/unique peaks between cCREs, pseudo-bulk called peaks, and TCGA-BRCA ATAC peaks. P-values were calculated using bedtools fisher (overlap significance). Fig. S10. DA-cCREs for each clusster. (A) Karyotype plot showing all DA-cCREs. Red indicates epiAneufinder-defined amplified regions. (B) Heatmap showing the motif enrichment for DA-cCREs for each cluster. The top 3 enriched homer known motifs for each set of DA-cCREs were displayed. (C) Heatmap showing the overlap significance between DA-cCREs and ER ChIP-seq distal peaks of primary breast tumors from Joosten et al., 2024 (its Supplementary Table 2). Fig. S11. pEPN construction (A) The flow of construction of pEPN. (B) Karyotype plot showing 136,594 cCREs with co-accessibility to the promoters, and epiAneufinder-defined amplified regions (red). (C) Dot plot showing the number of correlated cCREs overlapping with epiAneufinder-defined amplified regions. Red indicates the genes identified as amplified in breast cancer by a previous report (TCGA, Nature, 2012). (D) Dot plot showing the number of correlated cCREs per promoter. (E) Genome track plot showing CCND1 loci. The arcs and red box at the top of the track indicate correlated cCREs to the CCND1 promoter. The red box at the bottom indicates epiAneufinder-defined amplified regions. (F) Bar plot showing the distribution of the number of putative enhancers (pEnhancers). (G) Bar plot showing the distribution of the number of the promoters per pEnhancer. Fig. S12. Comparison between pEPN and Super-enhancer (SE) detection using ATAC-seq signals. (A) The flow of SE analysis using the peaks and insertions per cluster. (B) Bar plot showing the number of detected SEs per cluster. (C) Schema of target gene prediction. (D) Bar plot showing the number of the predicted genes per cluster. Fig. S13. Validation of our dataset and analysis using external scATAC-seq

datasets. (A) UMAP plots of the scMultiome dataset (scATAC + scRNA) from Terekhanova et al., colored by clusters (left) and by samples (right).(B) UMAP overlays showing marker gene expression levels in the Terekhanova et al. dataset.(C) UMAP plots of the scATAC-seq dataset from Regner et al., colored by clusters (left) and by samples (right).(D) UMAP overlays showing marker gene activity scores in the Regner et al. dataset.(E) Bar plot showing the proportion of hg38-converted cCREs overlapping with peaks identified in external datasets.(F) Heatmap showing the overlap between our DA-cCREs and cluster-specific peaks from the external datasets. The accompanying bar plot indicates the number of cluster-specific peaks.(G) Pie chart showing the proportion of genes in our pEPN analysis that share at least one putative enhancer with peak-to-gene linkages identified in external datasets. Fig. S14. Motif differential analysis between Ep34 and Ep35. Volcano plot showing the differential analysis result of the motif scores (Homer known motifs) between Ep34 and Ep35.

**Acknowledgements**
We thank Nami Nakasuji for her analytical support. We also thank Dr. Kenichi Miyata and all members of our laboratory for critical discussion and comments on the manuscripts. We also thank Enago (www.enago.jp) for the English language review.

**Authors' contributions**
CH, KK and RM performed data analysis. SS, YT, KO, CT, YO, NU, YH, TT, SM, and TU recruited patients and obtained clinical specimens. SS and LY processed tumour samples and performed ATAC-seq experiments. TO performed the pathological analysis. ToN provided support on experiments. TeN, SF, SO, TU and RM supervised all of the work. CH, KK and RM wrote the manuscript. All authors read and approved the final manuscript.

**Funding**
This work was supported in part by JSPS KAKENHI grant numbers JP20K07708 (to R. M.), JP24K02312 (to R. M.), JP24K18499 (to K. K.), and JP22K07161 (to To. N.), the Vehicle Racing Commemorative Foundation (to L. Y., R. M.), Project Mirai Cancer Research Grants (to K. K.), Mochida Memorial Foundation for Medical and Pharmaceutical Research (to K. K.), and The Yasuda Medical Foundation (to K. K).

**Data availability**
Processed scATAC-seq fragment data have been deposited at GEO (GSE299408) [30]. Raw sequencing data cannot be shared due to patient privacy concerns. The publicly available scATAC-seq data were downloaded from HTAN DCC portal (https://data.humantumoratlas.org) [49] under the HTAN WUSTL Atlas (Terekhanova et al.) and GEO accession number GSE243526 [50]. All original codes for reproducing the analyses are publicly available at [https://github.com/KoheiKumegawa/cCRE_JFCR-BRCA](https://github.com/KoheiKumegawa/cCRE_JFCR-BRCA).

## Declarations

**Ethics approval and consent to participate**
This study was conducted in accordance with the Declaration of Helsinki and its amendments. All participants provided written informed consent to participate prior to specimen collection. The protocol received approval from the institutional review board of the Cancer Institute Hospital, Japanese Foundation for Cancer Research (No. 2018-1168).

**Consent for publication**
All participants provided written informed consent for publication prior to specimen collection.

**Competing interests**
The authors declare no competing interests.

**Author details**
[1]Division of Cancer Epigenomics, Cancer Institute, Japanese Foundation for Cancer Research, Tokyo, Japan. [2]Department of Molecular Pathology, Yokohama City University Graduate School of Medicine, Yokohama, Japan. [3]Cancer Cell Diversity Project, NEXT-Ganken Program, Japanese Foundation for Cancer Research, Tokyo, Japan. [4]Breast Surgical Oncology, Breast Oncology Center, Cancer Institute Hospital, Japanese Foundation for Cancer Research, Tokyo, Japan. [5]Breast Medical Oncology, Breast Oncology Center, Cancer Institute Hospital, Japanese Foundation for Cancer Research, Tokyo, Japan. [6]Division of Pathology, Cancer Institute, Japanese Foundation for Cancer Research, Tokyo, Japan. [7]Division of Cancer Genomics, Cancer Precision Medicine Center, Japanese Foundation for Cancer Research, Tokyo, Japan. [8]Director's Room, Cancer Institute, Japanese Foundation for Cancer Research, Tokyo, Japan. [9]Breast Oncology Center, Cancer Institute Hospital, Japanese Foundation for Cancer Research, Tokyo, Japan.

Received: 13 November 2024 Accepted: 7 October 2025
Published online: 27 November 2025

**References**

1. Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, Ai R, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699–710. https://doi.org/10.1038/s41586-020-2493-4.
2. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30. https://doi.org/10.1038/nature14248.
3. Kim S, Wysocka J. Deciphering the multi-scale, quantitative cis-regulatory code. Mol Cell. 2023;83(3):373–92. https://doi.org/10.1016/j.molcel.2022.12.032.
4. Zu S, Li YE, Wang K, Armand EJ, Mamde S, Amaral ML, et al. Single-cell analysis of chromatin accessibility in the adult mouse brain. Nature. 2023;624:378–89. https://doi.org/10.1038/s41586-023-06824-9.
5. Ameen M, Sundaram L, Shen M, Banerjee A, Kundu S, Nair S, et al. Integrative single-cell analysis of cardiogenesis identifies developmental trajectories and non-coding mutations in congenital heart disease. Cell. 2022;185:4937-4953.e23. https://doi.org/10.1016/j.cell.2022.11.028.
6. Thomas ED, Timms AE, Giles S, Harkins-Perry S, Lyu P, Hoang T, et al. Cell-specific *cis*-regulatory elements and mechanisms of non-coding genetic disease in human retina and retinal organoids. Dev Cell. 2022;57:820-836.e6. https://doi.org/10.1016/j.devcel.2022.02.018.
7. Trevino AE, Müller F, Andersen J, Sundaram L, Kathiria A, Shcherbina A, et al. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. Cell. 2021;184:5053-5069.e23. https://doi.org/10.1016/j.cell.2021.07.039.
8. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020;578:112–21. https://doi.org/10.1038/s41586-019-1913-9.
9. Bushweller JH. Targeting transcription factors in cancer — from undruggable to reality. Nat Rev Cancer. 2019;19:611–24. https://doi.org/10.1038/s41568-019-0196-7.
10. Dentro SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. Cell. 2021;184:2239-2254.e39. https://doi.org/10.1016/j.cell.2021.03.009.
11. Koschmann C, Nunez FJ, Mendez F, Brosnan-Cashman JA, Meeker AK, Lowenstein PR, et al. Mutated chromatin regulatory factors as tumor drivers in cancer. Cancer Res. 2017;77:227–33. https://doi.org/10.1158/0008-5472.CAN-16-2301.
12. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science (1979). 2018;362. https://doi.org/10.1126/science.aav1898. American Association for the Advancement of Science.
13. Kumegawa K, Saeki S, Takahashi Y, Yang L, Osako T, Nakadai T, et al. Chromatin profile-based identification of a novel ER-positive breast cancer subgroup with reduced ER-responsive element accessibility. Br J Cancer. 2023;128:1208–22. https://doi.org/10.1038/s41416-023-02178-1.
14. Swanton C, Bernard E, Abbosh C, André F, Auwerx J, Balmain A, et al. Embracing cancer complexity: hallmarks of systemic disease. Cell. 2024;187:1589–616. https://doi.org/10.1016/j.cell.2024.02.009.
15. Preissl S, Gaulton KJ, Ren B. Characterizing cis-regulatory elements using single-cell epigenomics. Nat Rev Genet. 2023;24:21–43. https://doi.org/10.1038/s41576-022-00509-1.
16. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al. ArchR is a scalable software package for integrative single-cell chromatin

accessibility analysis. Nat Genet Nature Research. 2021;53:403–11. https://doi.org/10.1038/s41588-021-00790-6.
17. Becker WR, Nevins SA, Chen DC, Chiu R, Horning AM, Guha TK, et al. Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. Nat Genet. 2022;54:985–95. https://doi.org/10.1038/s41588-022-01088-x.
18. Guilhamon P, Chesnelong C, Kushida MM, Nikolic A, Singhal D, MacLeod G, et al. Single-cell chromatin accessibility profiling of glioblastoma identifies an invasive cancer stem cell population associated with lower survival. Elife. 2021. https://doi.org/10.7554/eLife.64090.
19. Russnes HG, Lingjærde OC, Børresen-Dale A-L, Caldas C. Breast cancer molecular stratification. Am J Pathol. 2017;187:2152–62. https://doi.org/10.1016/j.ajpath.2017.04.022.
20. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160–7. https://doi.org/10.1200/JCO.2008.18.1370.
21. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530–6. https://doi.org/10.1038/415530a.
22. Morganti S, Marra A, Crimini E, D'Amico P, Zagami P, Curigliano G. Refining risk stratification in HR-positive/HER2-negative early breast cancer: how to select patients for treatment escalation? Breast Cancer Res Treat. 2022;192:465–84. https://doi.org/10.1007/s10549-022-06535-1.
23. Bushnell GG, Deshmukh AP, den Hollander P, Luo M, Soundararajan R, Jia D, et al. Breast cancer dormancy: need for clinically relevant models to address current gaps in knowledge. NPJ Breast Cancer. 2021. https://doi.org/10.1038/s41523-021-00269-x.
24. Yu NY, Iftimi A, Yau C, Tobin NP, van 't Veer L, Hoadley KA, et al. Assessment of long-term distant recurrence-free survival associated with tamoxifen therapy in postmenopausal patients with luminal A or luminal B breast cancer. JAMA Oncol. 2019. https://doi.org/10.1001/jamaoncol.2019.1856.
25. Zhou Y, Li T, Choppavarapu L, Fang K, Lin S, Jin VX. Integration of scHi-C and scRNA-seq data defines distinct 3D-regulated and biological-context dependent cell subpopulations. Nat Commun. 2024;15:8310. https://doi.org/10.1038/s41467-024-52440-0.
26. Fang K, Ohihoin AG, Liu T, Choppavarapu L, Nosirov B, Wang Q, et al. Integrated single-cell analysis reveals distinct epigenetic-regulated cancer cell states and a heterogeneity-guided core signature in tamoxifen-resistant breast cancer. Genome Med. 2024;16:134. https://doi.org/10.1186/s13073-024-01407-3.
27. Iglesia MD, Jayasinghe RG, Chen S, Terekhanova NV, Herndon JM, Storrs E, et al. Differential chromatin accessibility and transcriptional dynamics define breast cancer subtypes and their lineages. Nat Cancer. 2024;5:1713–36. https://doi.org/10.1038/s43018-024-00773-6.
28. Regner MJ, Garcia-Recio S, Thennavan A, Wisniewska K, Mendez-Giraldez R, Felsheim B, et al. Defining the regulatory logic of breast cancer using single-cell epigenetic and transcriptome profiling. Cell Genomics. 2025;5:100765. https://doi.org/10.1016/j.xgen.2025.100765.
29. Kumegawa K, Takahashi Y, Saeki S, Yang L, Nakadai T, Osako T, et al. GRHL2 motif is associated with intratumor heterogeneity of cis-regulatory elements in luminal breast cancer. NPJ Breast Cancer. 2022;8:70. https://doi.org/10.1038/s41523-022-00438-6.
30. Hori C, Kumegawa K, Maruyama R. Characterizing heterogeneous cis-regulatory elements in gene regulatory programs associated with breast cancer. Gene Expression Omnibus. 2025. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE299408.
31. Lareau CA, Duarte FM, Chew JG, Kartha VK, Burkett ZD, Kohlway AS, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nat Biotechnol. 2019;37:916–24. https://doi.org/10.1038/s41587-019-0147-6.
32. Picelli S, Björklund ÅK, Reinius B, Sagasser S, Winberg G, Sandberg R. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. 2014;24:2033–40. https://doi.org/10.1101/gr.177881.114.
33. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with harmony. Nat Methods. 2019;16:1289–96. https://doi.org/10.1038/s41592-019-0619-0.
34. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184:3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.
35. Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, et al. Gene set knowledge discovery with Enrichr. Curr Protoc. 2021;1. https://doi.org/10.1002/cpz1.90.
36. Pal B, Chen Y, Vaillant F, Capaldo BD, Joyce R, Song X, et al. A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. EMBO J. 2021;40:e107333. https://doi.org/10.15252/embj.2020107333.
37. Ramakrishnan A, Symeonidi A, Hanel P, Schmid KT, Richter ML, Schubert M, et al. epiAneufinder identifies copy number alterations from single-cell ATAC-seq data. Nat Commun. 2023;14:5846. https://doi.org/10.1038/s41467-023-41076-1.
38. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nat Protoc. 2012;7:1728–40. https://doi.org/10.1038/nprot.2012.101.
39. Yu G, Wang L-G, He Q-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics. 2015;31:2382–3. https://doi.org/10.1093/bioinformatics/btv145.
40. Yang L, Kumegawa K, Saeki S, Nakadai T, Maruyama R. Identification of lineage-specific epigenetic regulators FOXA1 and GRHL2 through chromatin accessibility profiling in breast cancer cell lines. Cancer Gene Ther. 2024. https://doi.org/10.1038/s41417-024-00745-z.
41. Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell. 2013;153:320–34. https://doi.org/10.1016/j.cell.2013.03.036.
42. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;153:307–19. https://doi.org/10.1016/j.cell.2013.03.035.
43. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013.
44. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9. https://doi.org/10.1093/bioinformatics/btp352.
45. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. PLoS Comput Biol. 2016;12:e1004873. https://doi.org/10.1371/journal.pcbi.1004873.
46. Deshpande V, Luebeck J, Nguyen N-PD, Bakhtiari M, Turner KM, Schwab R, et al. Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. Nat Commun. 2019;10:392. https://doi.org/10.1038/s41467-018-08200-y.
47. Luebeck J, Ng AWT, Galipeau PC, Li X, Sanchez CA, Katz-Summercorn AC, et al. Extrachromosomal DNA in the cancerous transformation of Barrett's oesophagus. Nature. 2023;616:798–805. https://doi.org/10.1038/s41586-023-05937-5.
48. Terekhanova NV, Karpova A, Liang W-W, Strzalkowski A, Chen S, Li Y, et al. Epigenetic regulation during cancer transitions across 11 tumour types. Nature. 2023;623:432–41. https://doi.org/10.1038/s41586-023-06682-5.
49. Human Tumor Atlas Network Portal Data Release V6.3. Human tumor atlas network. 2025. https://data.humantumoratlas.org.
50. Regner MJ, Franco HL. Defining the regulatory logic of breast cancer using single-cell epigenetic and transcriptome profiling. Gene Expression Omnibus. 2024. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE243526.
51. Venteicher AS, Tirosh I, Hebert C, Yizhak K, Neftel C, Filbin MG, et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science. 2017. https://doi.org/10.1126/science.aai8478.
52. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature. 2016. https://doi.org/10.1038/nature20123.
53. Filbin MG, Tirosh I, Hovestadt V, Shaw ML, Escalante LE, Mathewson ND, et al. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. Science. 2018. https://doi.org/10.1126/science.aao4750.
54. Joosten SEP, Gregoricchio S, Stelloo S, Yapıcı E, Huang C-CF, Yavuz K, et al. Estrogen receptor 1 chromatin profiling in human breast tumors reveals high inter-patient heterogeneity with enrichment of risk SNPs and enhancer activity at most-conserved regions. Genome Res. 2024. https://doi.org/10.1101/gr.278680.123.

55. Ito K, Park SH, Katsyv I, Zhang W, De Angelis C, Schiff R, et al. Ptk6 regulates growth and survival of endocrine therapy-resistant ER+ breast cancer cells. NPJ Breast Cancer. 2017;3:45. https://doi.org/10.1038/s41523-017-0047-1.
56. Chen Y, Qu W, Tu J, Yang L, Gui X. Prognostic impact of PTK6 expression in triple negative breast cancer. BMC Womens Health. 2023;23:575. https://doi.org/10.1186/s12905-023-02736-y.
57. Sundaram L, Kumar A, Zatzman M, Salcedo A, Ravindra N, Shams S, et al. Single-cell chromatin accessibility reveals malignant regulatory programs in primary human cancers. Science. 2024. https://doi.org/10.1126/science.adk9217.

## Publisher's Note