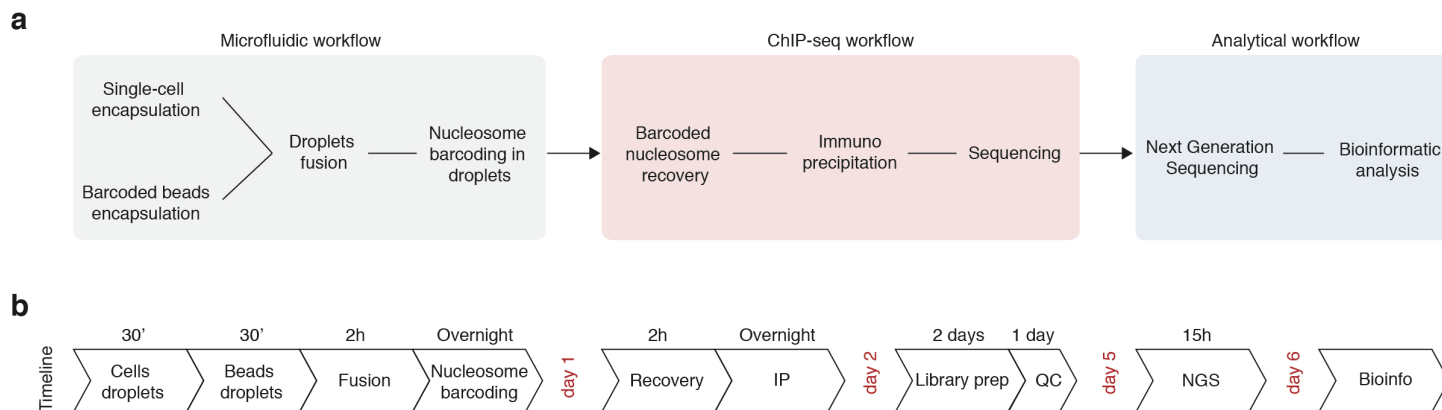


In the format provided by the authors and unedited.

High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer

Kevin Grosselin ^{1,2,9}, Adeline Durand ^{3,4}, Justine Marsolier^{3,4}, Adeline Poitou^{1,10},
Elisabetta Marangoni⁴, Fariba Nemati⁴, Ahmed Dahmani⁴, Sonia Lameiras⁵, Fabien Reyat^{4,6,7},
Olivia Frenoy^{1,11}, Yannick Pousse¹, Marcel Reichen^{1,12}, Adam Woolfe¹, Colin Brenan^{1,8},
Andrew D. Griffiths ^{2,13*}, Céline Vallot ^{3,4,13*} and Annabelle Gérard ^{1,13*}

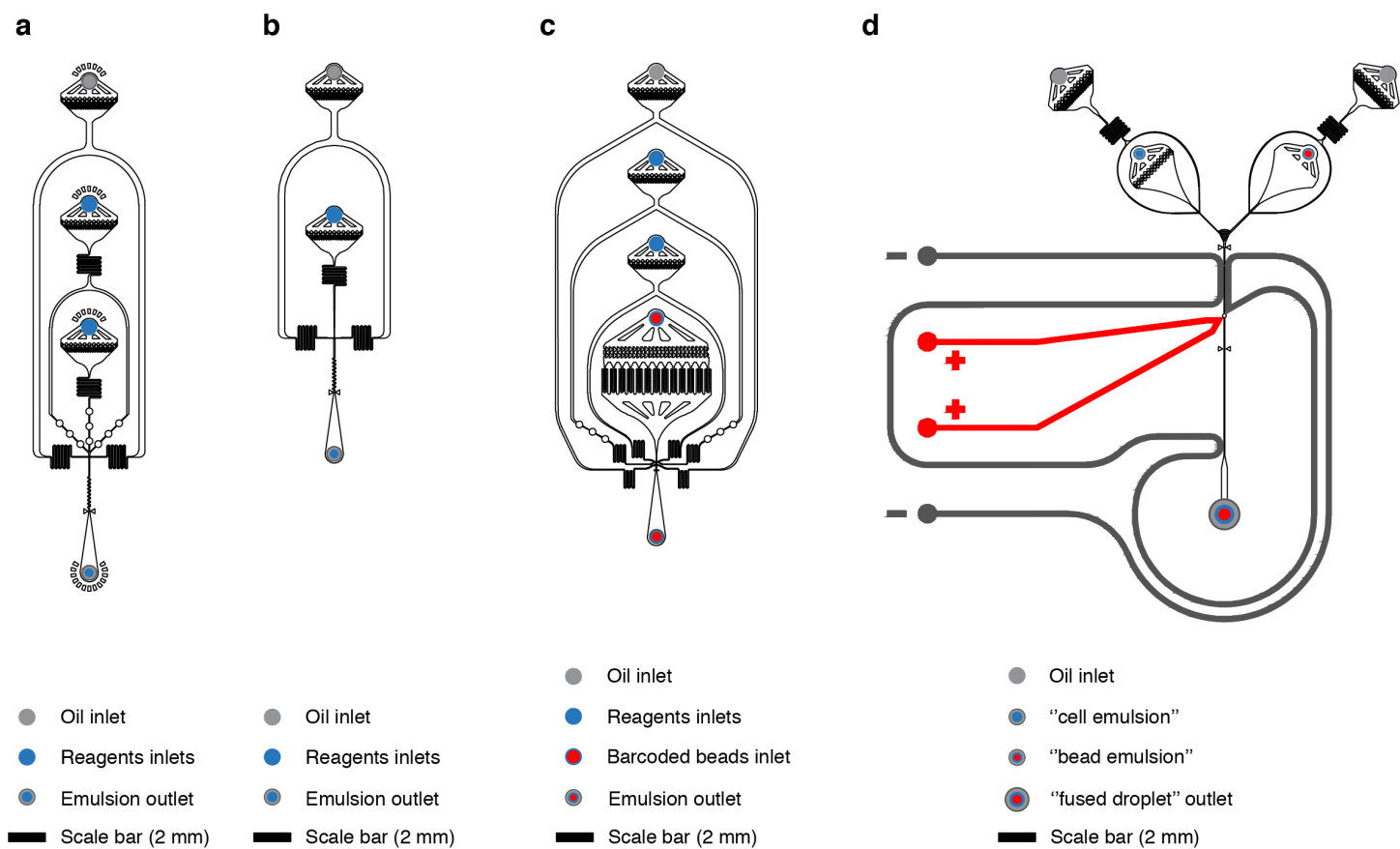
¹HiFiBio SAS, Paris, France. ²Laboratoire de Biochimie, CNRS UMR8231, ESPCI Paris, PSL University, Paris, France. ³CNRS UMR3244, Institut Curie, PSL University, Paris, France. ⁴Translational Research Department Institut Curie, PSL University, Paris, France. ⁵Institut Curie Genomics of Excellence Platform, PSL University, Paris, France. ⁶INSERM U932, Institut Curie, PSL University, Paris, France. ⁷Surgical Department, Institut Curie, Paris, France. ⁸HiFiBio Inc., Cambridge, MA, USA. ⁹Present address: Broad Institute of MIT and Harvard, Cambridge, MA, USA. ¹⁰Present address: Unicancer, Paris, France. ¹¹Present address: Institut Pasteur, Paris, France. ¹²Present address: Bayer AG, Leverkusen, Germany. ¹³These authors jointly supervised this work: Andrew D. Griffiths, Céline Vallot, Annabelle Gérard. *e-mail: andrew.griffiths@espci.fr; celine.vallot@curie.fr; a.gerard@hifibio.com



Supplementary Figure 1

Overview of the single-cell ChIP-seq procedure

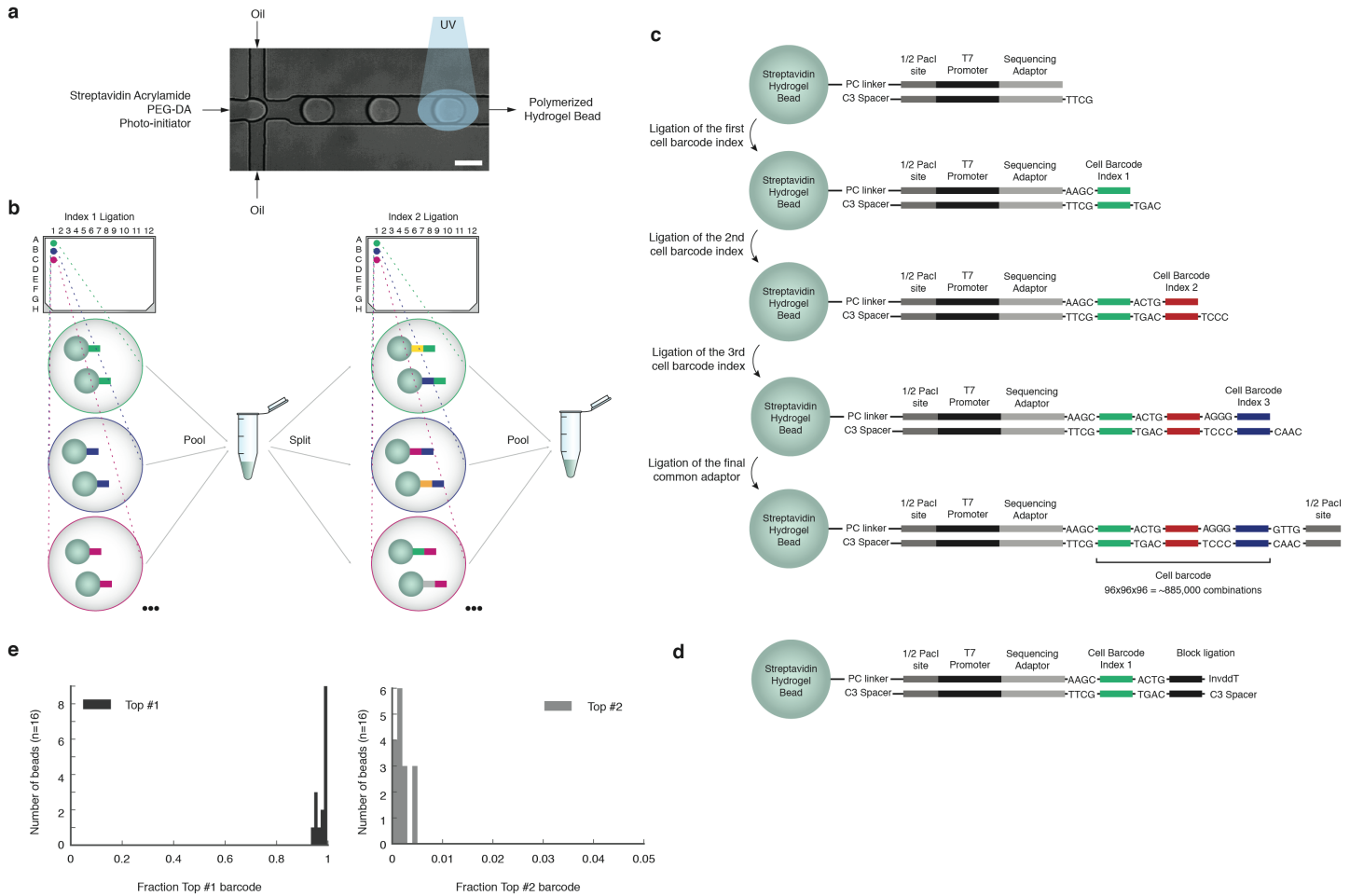
(a) The single-cell ChIP-seq procedure is divided into 3 workflows including a droplet-microfluidic workflow, a ChIP-seq workflow and an analytical workflow. Key steps of each workflow are indicated in the boxes. (b) Single-cell ChIP-seq timeline: the entire process from the cell harvesting to the NGS takes 5 days. IP: immunoprecipitation; QC: Quality Controls; NGS: Next Generation Sequencing.



Supplementary Figure 2

Design of microfluidic devices

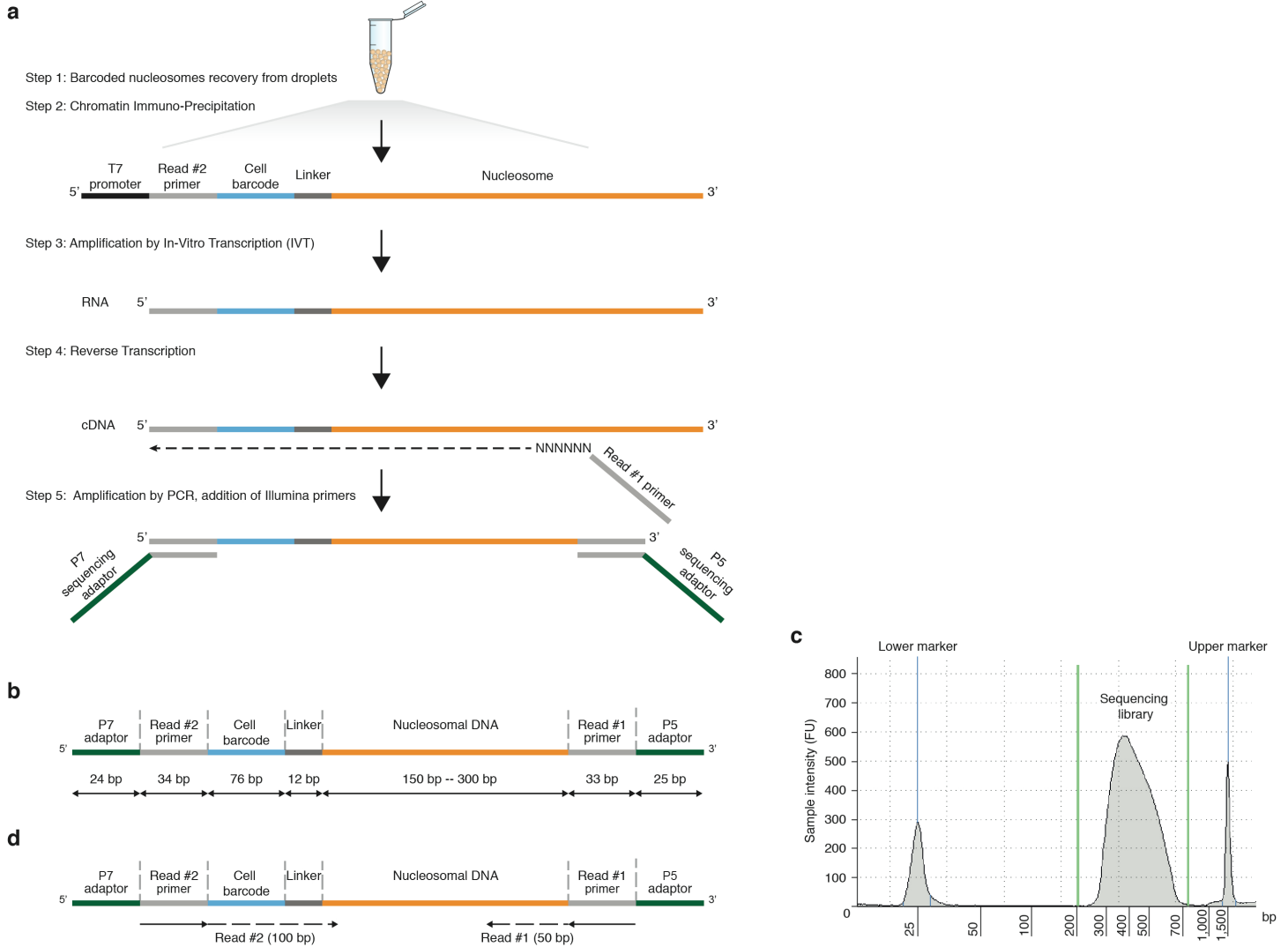
(a) Device used to compartmentalize cells in 45 μ l droplets. (b) Device used to produce 9 μ l droplets to make hydrogel beads. (c) Device used for compartmentalization of hydrogel beads in 100 μ l droplets. (d) Device used to merge nucleosome-containing droplets with hydrogel barcoded bead-containing droplets. Potential and ground electrodes are indicated with the '+' and the '-' marks, respectively. Scale bars correspond to 2 mm.



Supplementary Figure 3

Barcoded bead production and quality control

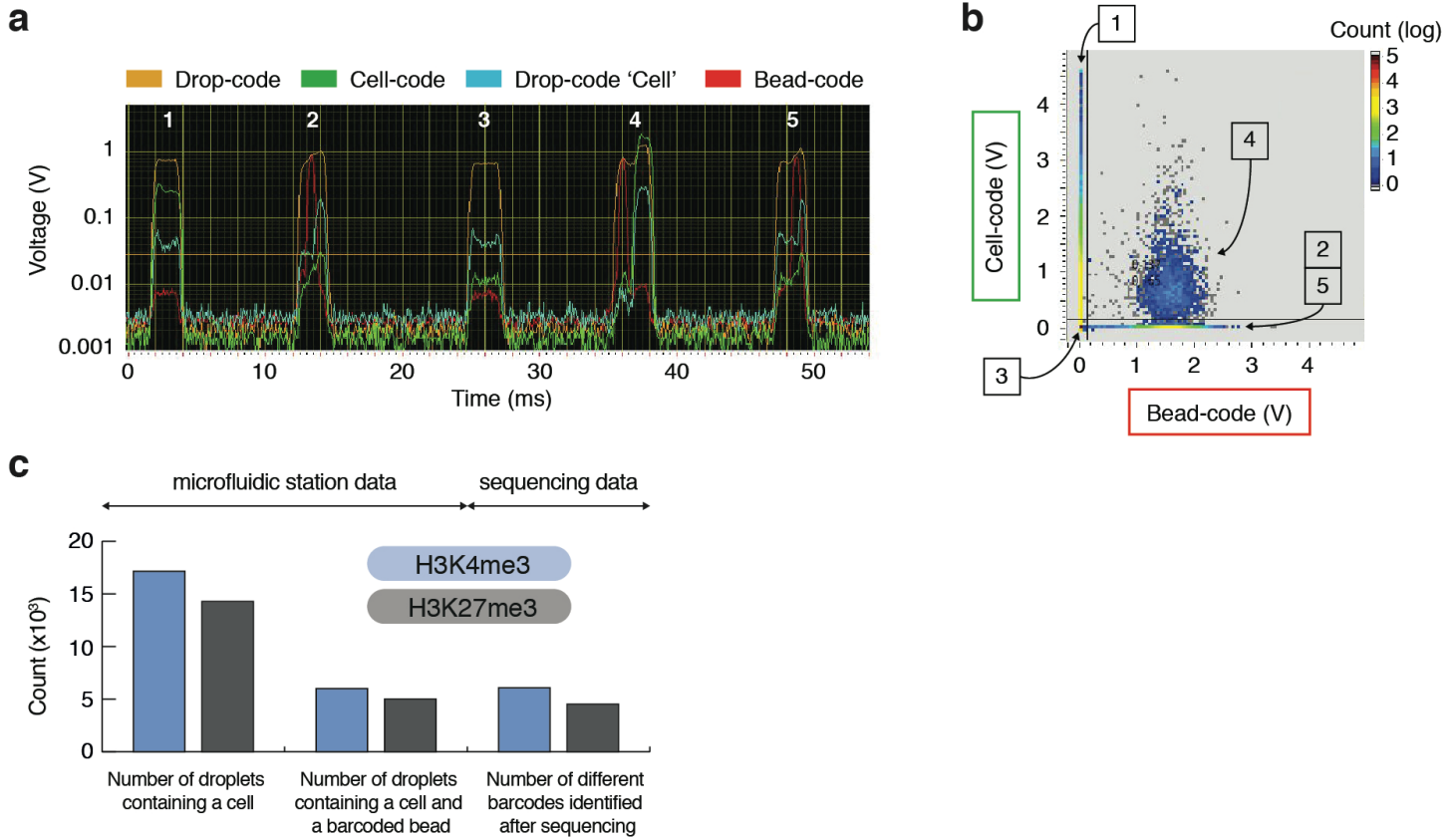
(a) Beads were produced in a microfluidic device with 2-inlets by dispersing a mixture comprising PolyEthylene Glycol Di-Acrydrite (PEG-DA), Streptavidin Acrylamide and the photo-initiator. Flow rates were adjusted to produce 9 μ l droplets, and immediately exposed to UV light for polymerization of the hydrogel network (see Methods). Scale bar corresponds to 25 μ m. (b) Split- and-pool synthesis principle for the addition of successive indexes. (c) Barcodes were synthesized by successive ligation of double-stranded indexes containing 5' overhang of 4 base pairs by three rounds of split-and-pool synthesis using 96 Index 1, 96 Index 2 and 96 Index 3. Barcodes were flanked at one end by common sequences comprising a 1/2 Pac1 restriction site, a T7 promoter and the Illumina Read #2 sequencing primer, which were bound to the beads via a photocleavable linker (PC-linker). A 3' C3-spacer was added to the 3'end of the photocleaved site for directed ligation to the other end of the barcode comprising a second common sequence with the 1/2 Pac1 restriction site ligated to the index 3. (d) Barcodes that failed in one of the three split-pool rounds were completed with a "block" oligonucleotide comprising a 5' C3-spacer and a 3' Inverted ddT to prevent ligation. (e) Single-bead sequencing results showing the fraction of the first two most abundant barcodes of 16 beads. In average, 97.7% of the barcodes corresponded to the same sequence and the second most abundant barcode represented 0.17% of all sequences.



Supplementary Figure 4

Sequencing library preparation

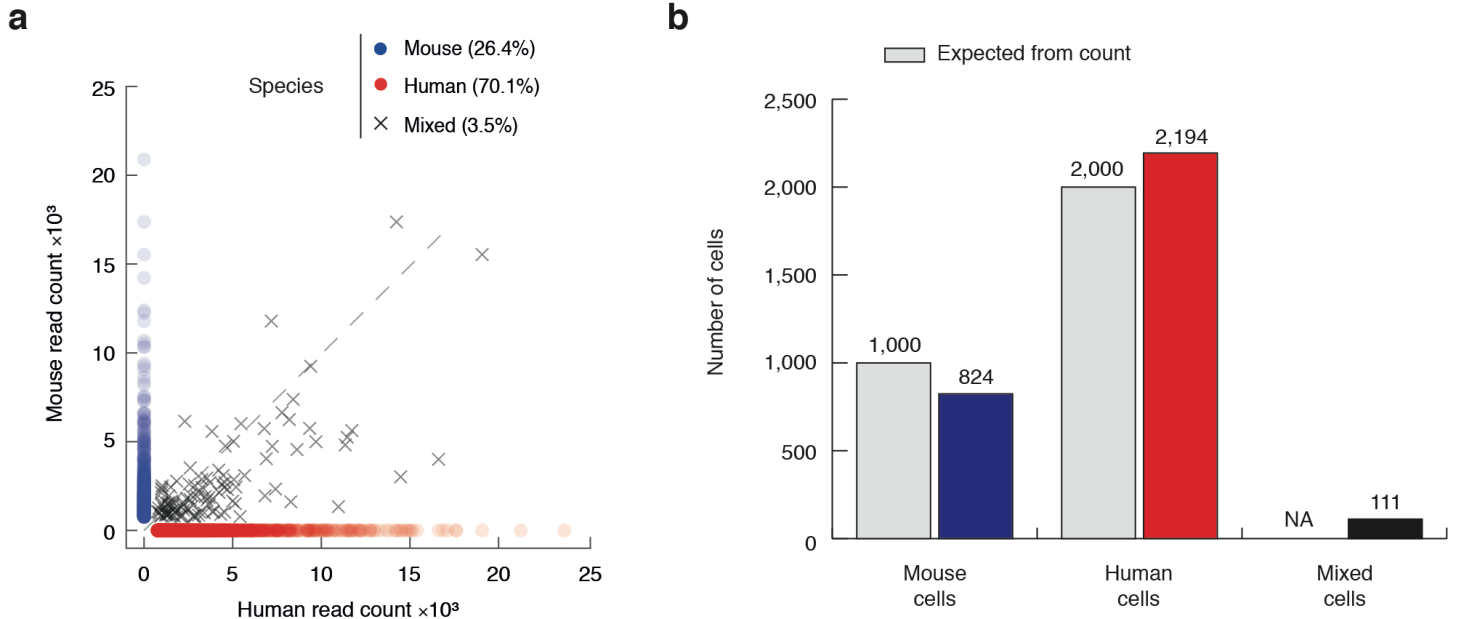
(a) Enriched barcoded nucleosomes were linearly amplified by *in vitro* transcription. The amplified RNAs were reverse transcribed into cDNA by random priming, appending a reverse complement of Illumina Read #1 sequencing primer. The cDNAs were amplified by PCR, appending an Illumina P7 and P5 sequences. (b) Schematic of the final sequencing product with size in bp of each element constituting the sequence. (c) Electropherogram showing the size distribution of the final sequencing library post agarose gel purification. The smear ranges from 300 bp to 700 bp and corresponds to barcoded nucleosomes (profile obtained by TapeStation [Agilent]). (d) Single-cell ChIP-seq libraries were sequenced as follows: 50 bp were assigned to read the nucleosomal sequence and 100 bp were assigned to read the barcode.



Supplementary Figure 5

Live monitoring of the fusion between droplets with cells and with barcoded hydrogel beads on the microfluidic system

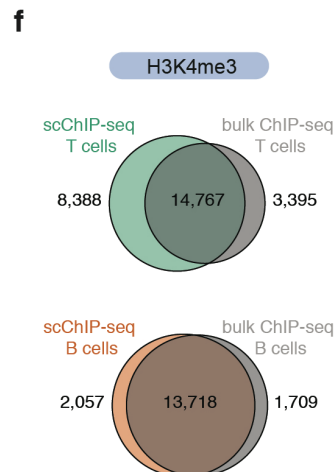
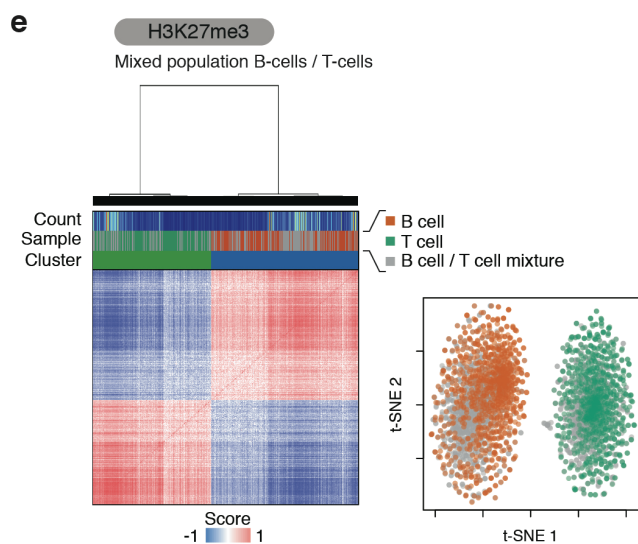
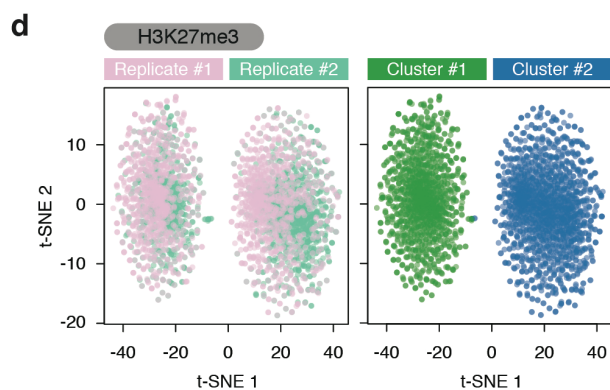
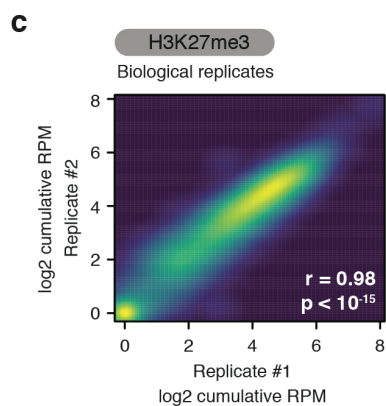
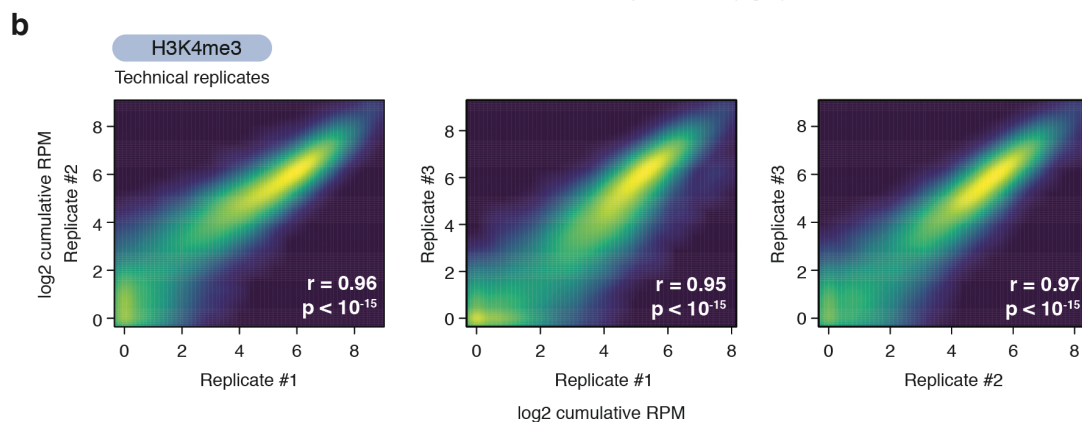
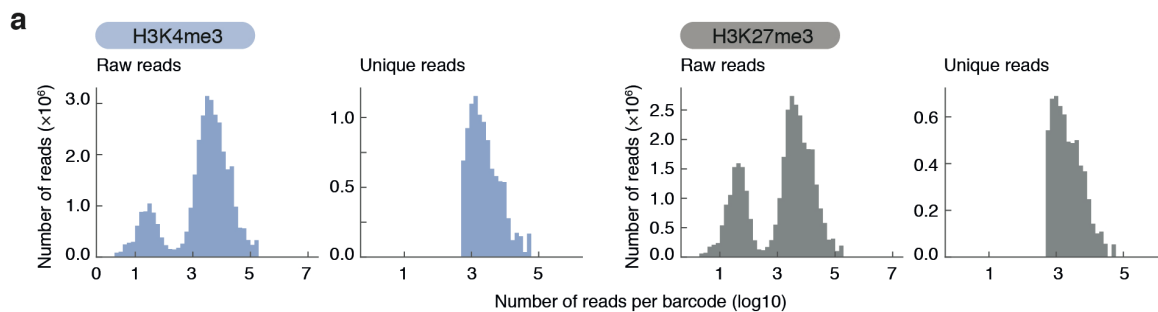
(a) Time traces of 5 droplets showing possible fusion events occurring in all scChIP-seq experiments: droplet #1 contained a cell but no bead; droplet #2 and droplet #5 contained one bead but no cell; droplet #3 is empty; droplet #4 contained one cell plus one bead, resulting in nucleosomes barcoding. Droplets were scanned after fusion as they crossed a laser beam and their fluorescence intensity analyzed in real time. Sulforhodamine B (orange fluorescent) was used as common marker (drop-code) and Dye-405 (violet fluorescent) was used as marker of the nucleosome-containing droplets (drop-code 'Cell'). Cells and beads were labeled with Calcein AM cell-permeant dye (green fluorescent) and Biotin-Cy5 (red fluorescent, 'Bead-code'), respectively. (b) Scatter plot showing cell fluorescence intensity (green, 'Cell-code') versus bead fluorescence intensity (red, 'Bead-code') in each droplet allowing precise counting of the number of cells co-encapsulated with a barcoded bead. Droplets from panel (a) are indicated as examples of the different merged droplet populations. (c) Barplot showing for two scChIP-seq experiments targeting distinct histone modifications (H3K27me3 and H3K4me3), the total number of encapsulated cells detected by fluorescence on the microfluidic station, the number of cells co-encapsulated with a barcoded hydrogel bead detected by fluorescence on the microfluidic station and the number of cells (*i.e.* unique barcodes) identified from the sequencing data.



Supplementary Figure 6

Human and mouse cells mixture confirms single-cell resolution

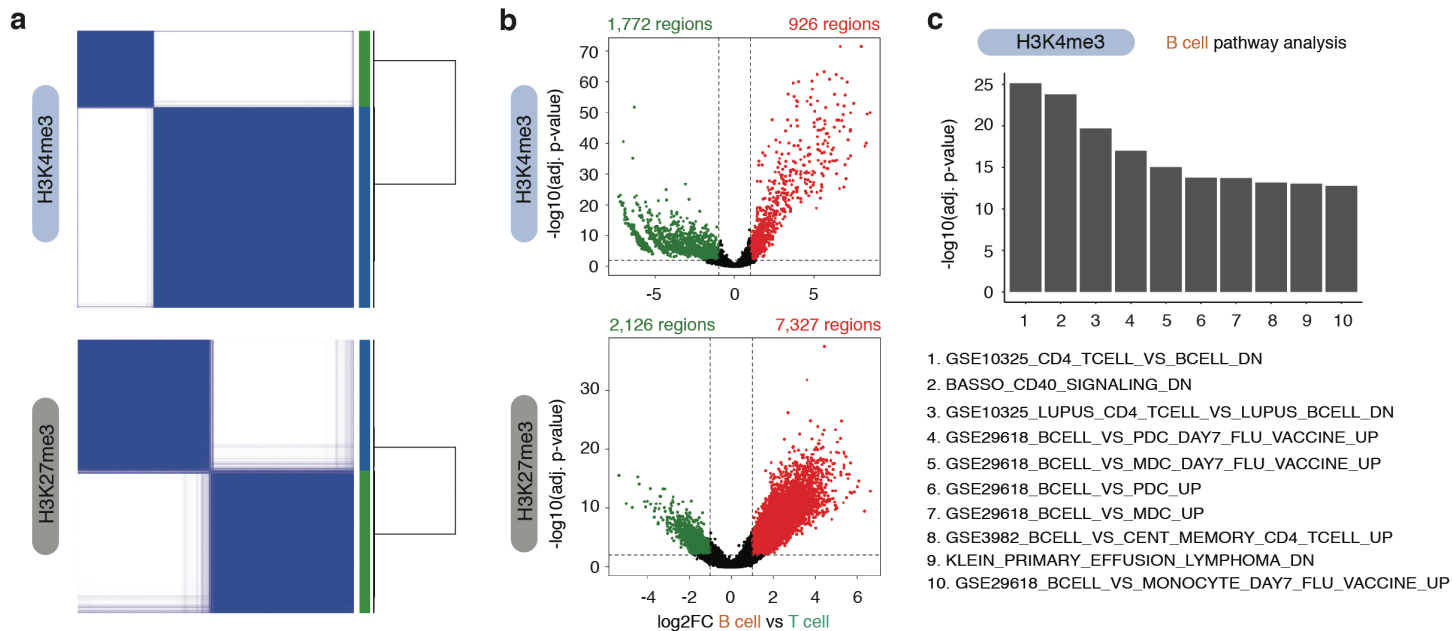
(a) Scatter plot of number of reads per barcode aligning to the mouse versus human reference genome showing that 96.5% of the barcodes are specific to one species (at least 95% of the reads with the same barcode mapping to one of the two species). The percentage of mouse (26.4%), human (70.1%) and mixed (3.5%) species was close to expected values based on a Poisson distribution of cells in droplets with an average number of cells per droplets λ of 0.1 (32.6%, 65.2% and 2.2%, respectively). (b) Barplot showing the number of barcodes (corresponding to the number of cells) identified for each species (blue, red and black bars) in comparison to the expected number of cells counted on the microfluidic station (gray bars; 3,000 in total from a mixture comprising 1/3 mouse cells and 2/3 of human cells).



Supplementary Figure 7

H3K4me3 and H3K27me3 single-cell ChIP-seq datasets for *in vitro* experiments

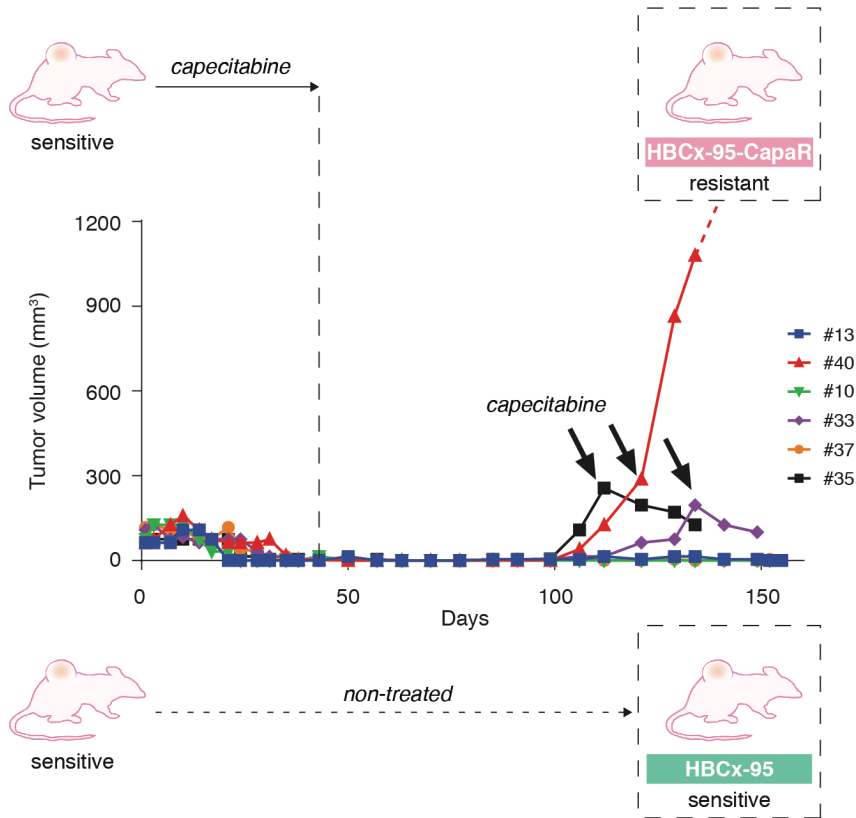
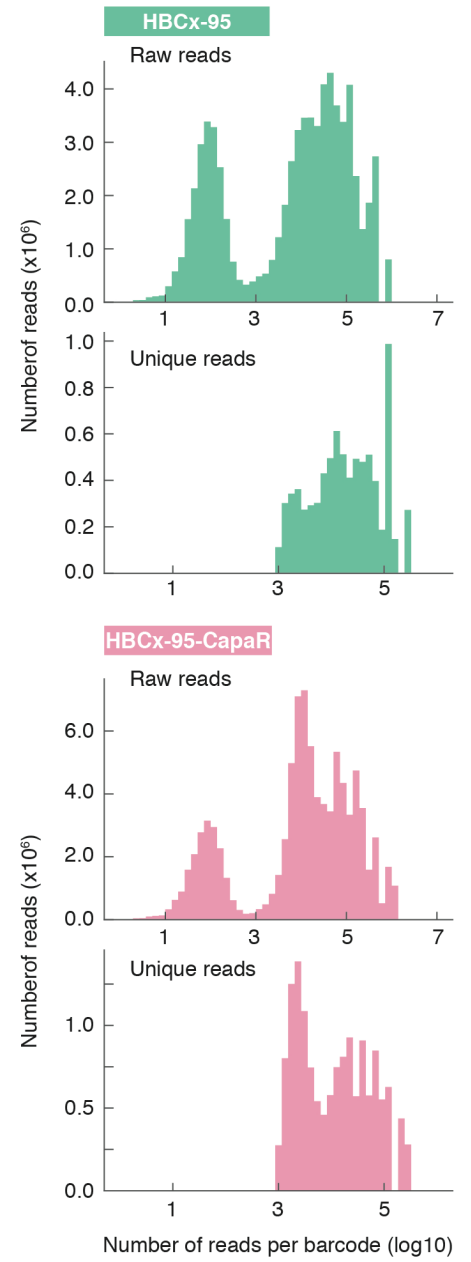
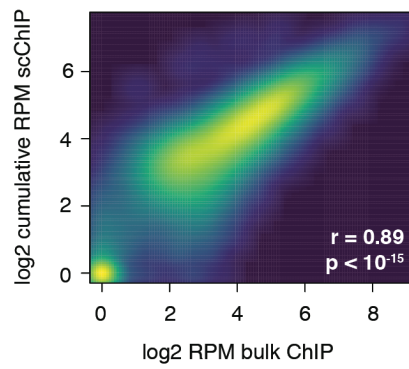
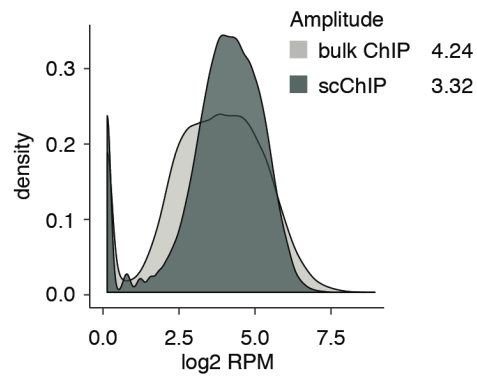
(a) Histograms of the distribution of scChIP-seq raw and unique sequencing reads per barcode (i.e. per cell) in H3K4me3 and H3K27me3 single-cell ChIP-seq datasets. (b) Density scatterplots representing log2 cumulative counts between three independent fractions of the same emulsion of B cells collected and processed in parallel to generate H3K4me3 single-cell ChIP-seq datasets. Correlation between replicates is calculated based on the cumulative count per million reads in 5 kb genomic bins across single-cells ($n = 642,098$ genomic bins). Pearson's correlation scores and P values are computed genome-wide. (c) Density scatter plot for two biological replicates corresponding to two emulsions of B cells collected from different cell culture flasks and processed with different batches of barcoded hydrogel beads to generate H3K27me3 single-cell ChIP-seq datasets. Correlation between replicates is calculated based on the cumulative count per million reads in 50 kb genomic bins across single-cells ($n = 64,455$ genomic bins). Pearson's correlation score and P value are computed genome-wide. (d) t-SNE plots representing H3K27me3 scChIP-seq data from two biological replicates ($n = 1,813$ and $1,451$ single-cells in replicate #1 and #2 respectively), colored according to the batch of origin (left) or consensus clustering results (right), arguing against a batch effect for cell population clustering. (e) Left panel: hierarchical clustering and corresponding heatmap of cell-to-cell Pearson's correlation scores for H3K27me3 scChIP-seq data from a mixed population of 1:1 human B cells and T cells together ($n = 1,453$ single-cells) with separately barcoded B cells ($n = 922$ single-cells) and T cells ($n = 879$ single-cells). The unique read count, the batch of origin and the consensus clustering result is indicated above the heatmap. Right panel: Corresponding t-SNE plot, points from the mixed population are colored in gray, while points from the separately barcoded B cells and T cells are colored according to the cell type-specific barcode sequence. (f) Venn diagrams comparing for T and B cells datasets the H3K4me3 peaks detected by single-cell and bulk approaches.



Supplementary Figure 8

Single-cell ChIP-seq data distinguish human T cells (Jurkat) from human B cells (Ramos)

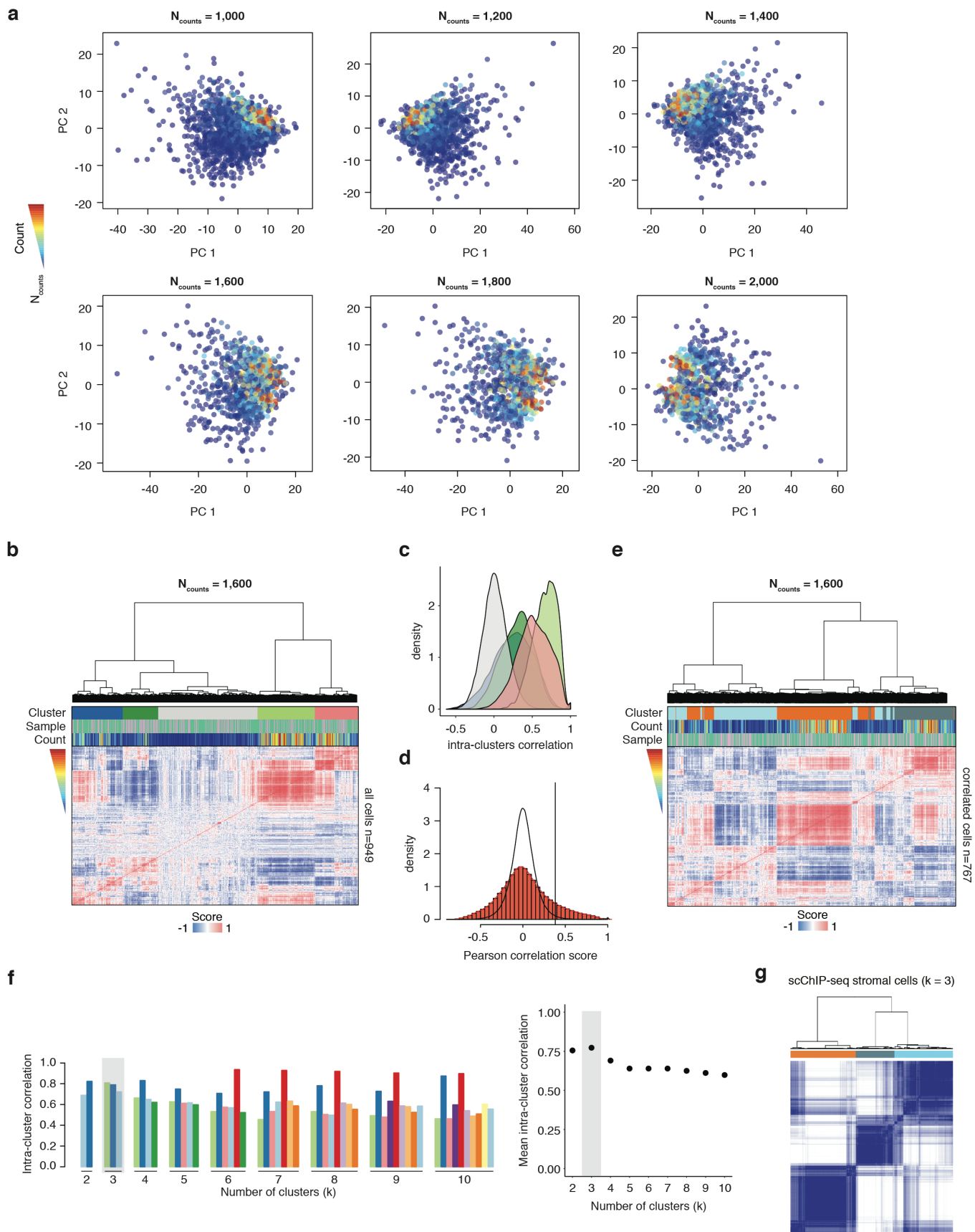
(a) Consensus clustering matrix for H3K4me3 (top panel) and H3K27me3 (bottom panel) scChIP-seq datasets. Consensus score ranges from 0 (white: never clustered together) to 1 (dark blue: always clustered together). (b) Volcano plots representing adjusted P values (two-sided Wilcoxon signed-rank test) versus fold-changes for differential analysis comparing chromatin features between B cells and T cells (thresholds of 0.01 for adjusted P value and 1 for $|\log_2FC|$) for H3K4me3 (top panel, $n = 1,772$ enriched and 926 depleted regions) and H3K27me3 (bottom panel, $n = 2,126$ enriched and 7,327 depleted regions) scChIP-seq datasets. (c) Barplot displaying the $-\log_{10}$ of adjusted P values from pathway analysis in H3K4me3 scChIP-seq dataset. The top 10 significant gene sets are indicated below the barplot.

a**b****c****d**

Supplementary Figure 9

Residual triple-negative breast tumor model of acquired resistance to Capecitabine

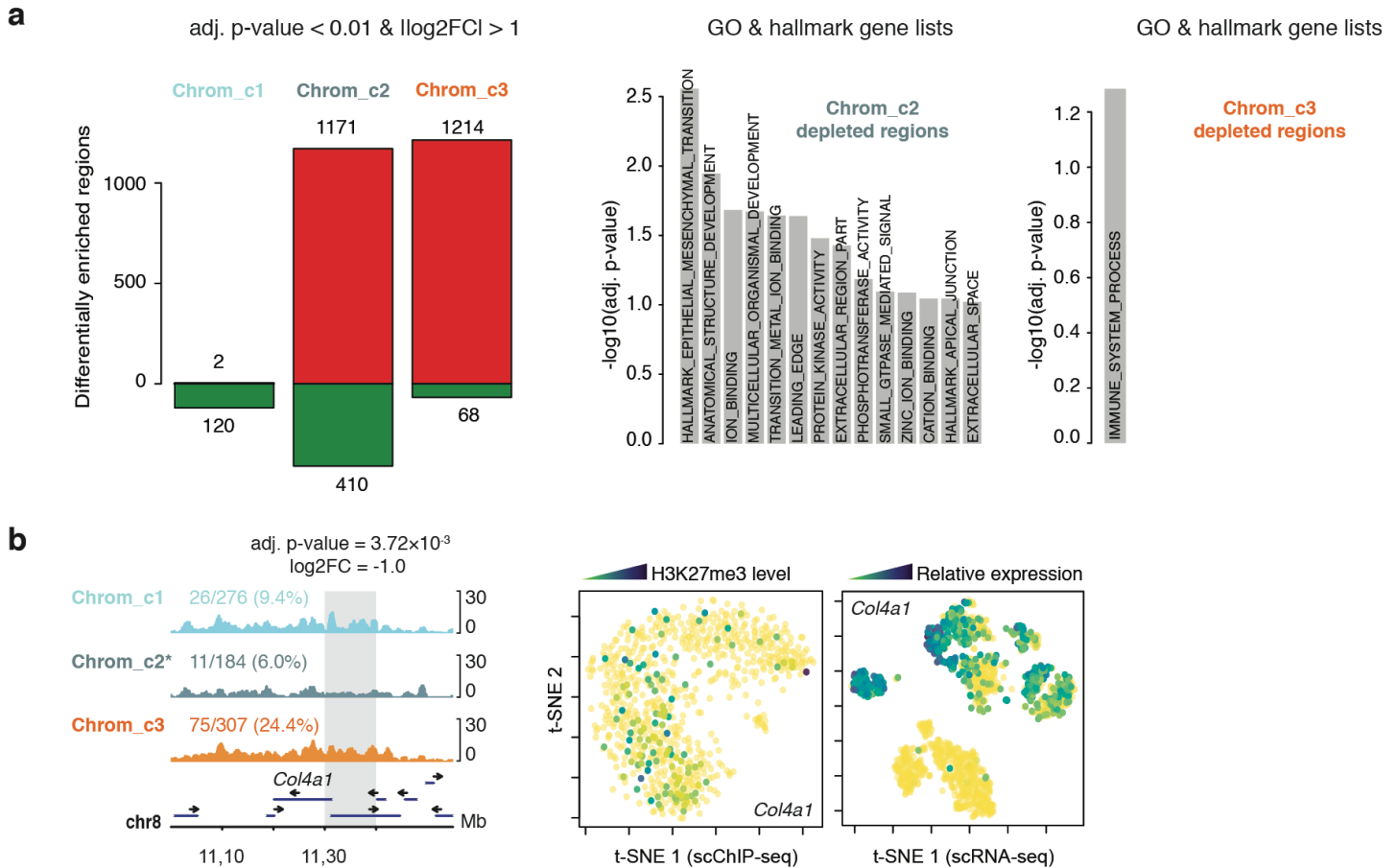
(a) $n = 6$ mice with HBCx-95 derivatives were treated with Capecitabine for 6 weeks. Graphs represent relative tumor volumes (RTV, mm^3) with time. Mice with recurrent tumors were treated for a second round of Capecitabine when PDX reached a volume of over 200 mm^3 (mice #35, #40 & #33). Mouse #40 did not respond to Capecitabine and the PDX specimen was extracted at 1080 mm^3 and tagged as HBCx-95-CapaR. Samples studied in manuscript are indicated with a dashed rectangle. (b) Histograms of the distribution of scChIP-seq raw and unique sequencing reads per cell in untreated HBCx-95 and Capecitabine-resistant HBCx-95-CapaR PDX. (c) Scatter plot displaying the $\log_2$ enrichments in cumulative single-cell versus bulk H3K27me3 ChIP-seq data, calculated based on the counts per million mapped reads in 50 kb genomic bins ($n = 54,650$ genomic bins). Pearson's correlation score and P value are computed genome-wide. (d) Distribution of the $\log_2$ enrichments in 50 kb genomic bins for cumulative single-cell and bulk H3K27me3 ChIP-seq data. Amplitude between maximum and minimum $\log_2$ enrichments are indicated on the top right for each ChIP-seq method (either bulk or single-cell).



Supplementary Figure 10

Detailed unsupervised analysis of scChIP-seq H3K27me3 profiles for mouse stromal cells from the HBCx-95 model

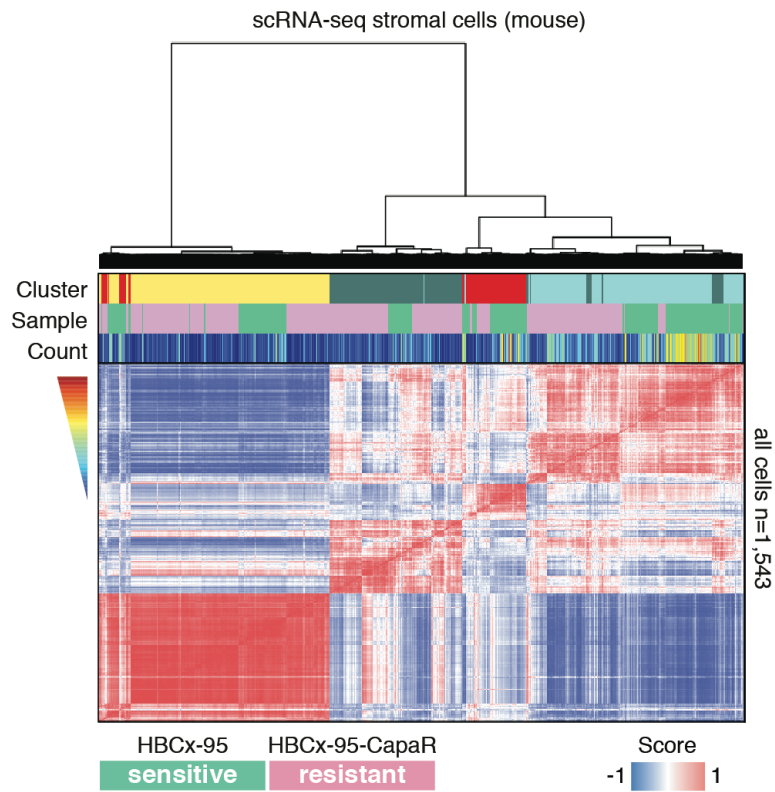
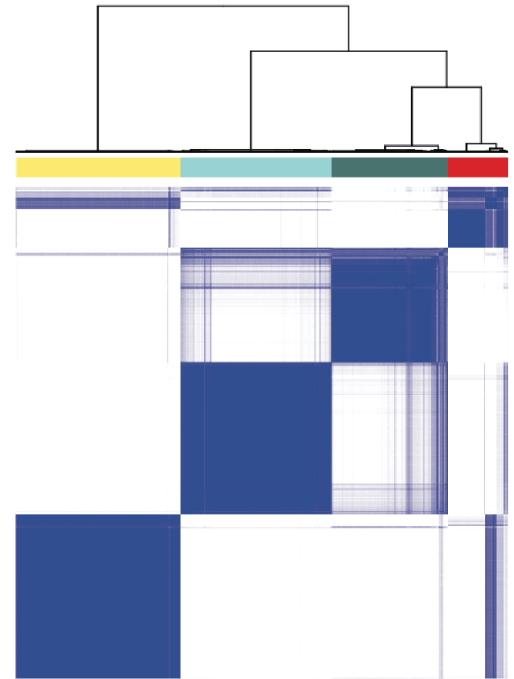
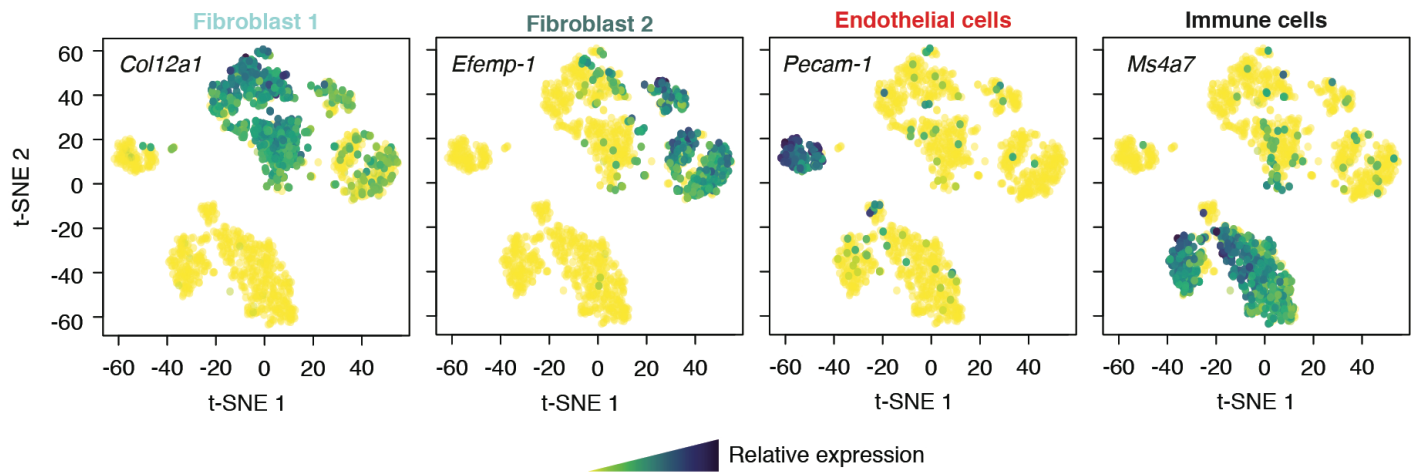
(a) PCA plots with varying minimum number of unique mapped reads per cell showing that H3K27me3 scChIPseq principal components were mainly driven by cell coverage up to 1,600 reads per cell. (b) Hierarchical clustering and corresponding heatmap of cell-to-cell Pearson's correlation scores based on their H3K27me3 profiles for $n=949$ mouse stromal cells with a minimum coverage of 1,600 unique mapped reads. Cluster membership from hierarchical clustering, sample of origin (green for HBCx-95 and pink for HBCx-95-CapaR) and unique read count are indicated above the heatmap. (c) Distribution of the intra-cluster correlation scores for each cluster identified by hierarchical clustering. (d) Distribution of cell-to-cell Pearson's correlation scores for real scChIP-seq datasets (in red) and randomized datasets (black curve). The upper percentile is indicated by a vertical black line. (e) Hierarchical clustering and corresponding heatmap of cell-to-cell Pearson's correlation scores for 'correlated' cells only, representing 81.8% of all cells (see Methods). (f) Left: mean of all pairwise correlation scores between cluster's members is plotted for k clusters ranging from 2 to 10. Right: mean intra-cluster correlation score for k clusters ranging from 2 to 10. At $k = 3$ clusters, the intra-cluster correlation is maximized. (g) Hierarchical clustering and corresponding heatmap of cell-to-cell consensus clustering score for scChIP-seq mouse stromal cells. Consensus score ranges from 0 (white: never clustered together) to 1 (dark blue: always clustered together). Cluster membership is color coded above the heatmap.



Supplementary Figure 11

Differential analysis of single-cell ChIP-seq H3K27me3 profiles of mouse stromal cells from the HBCx-95 model

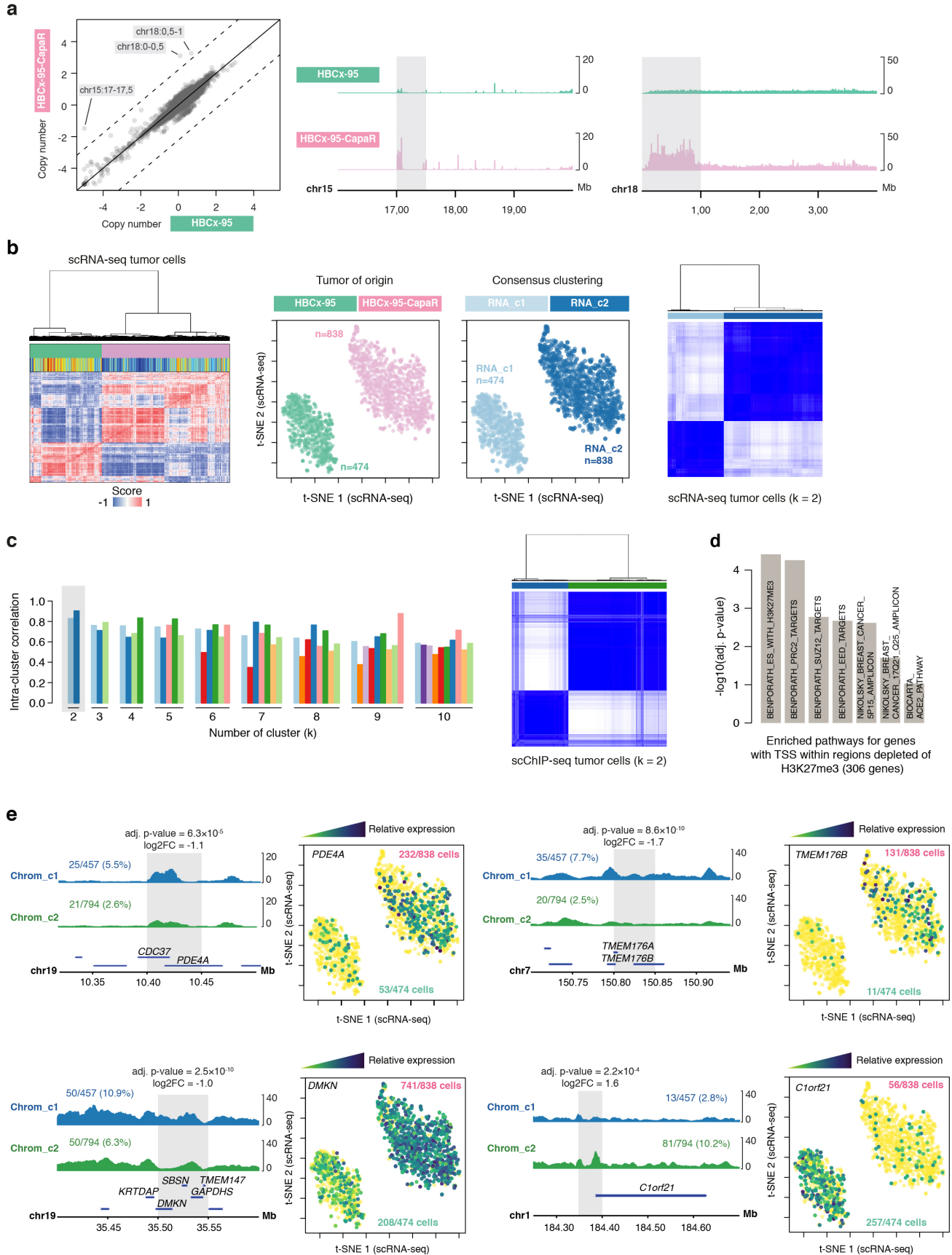
(a) Left panel: differentially bound regions in cells from each cluster versus all remaining cells identified by two-sided Wilcoxon signed-rank test. Genomic regions were considered enriched (red) or depleted (green) if the adjusted P values were lower than 0.01 and the absolute $\log_2$ fold change greater than 1. Middle and right panels: barplots displaying the $-\log_{10}$ of adjusted P values from pathway analysis (using MSigDB, adjusted P value < 0.1) for regions depleted in H3K27me3 specifically in *Chrom_c2* and *Chrom_c3*. The top 10 significant gene sets are indicated on the barplots. (b) Left panel: snapshot of a differentially depleted locus for cells from *Chrom_c2* (*Col4a1*) versus all other cells, with cumulative single-cell profiles for each cell type and cluster ($n = 276$, 184 and 307 single-cells for *Chrom_c1*, *Chrom_c2* and *Chrom_c3* respectively). Differentially bound regions identified by two-sided Wilcoxon signed-rank test is indicated in gray with the corresponding adjusted P value and $\log_2$ fold change, calculated in respect to *Chrom_c2*. The number and the proportion of cells with H3K27me3 enrichment within each cluster are indicated above tracks. Right panels: t-SNE representation of scChIP-seq and scRNA-seq datasets, dots are colored according to H3K27me3 enrichment and expression signals in each cell for *Col4a1* locus.

a**b****c**

Supplementary Figure 12

Clustering and differential analysis of single-cell RNA-seq profiles of mouse stromal cells from the HBCx-95 model

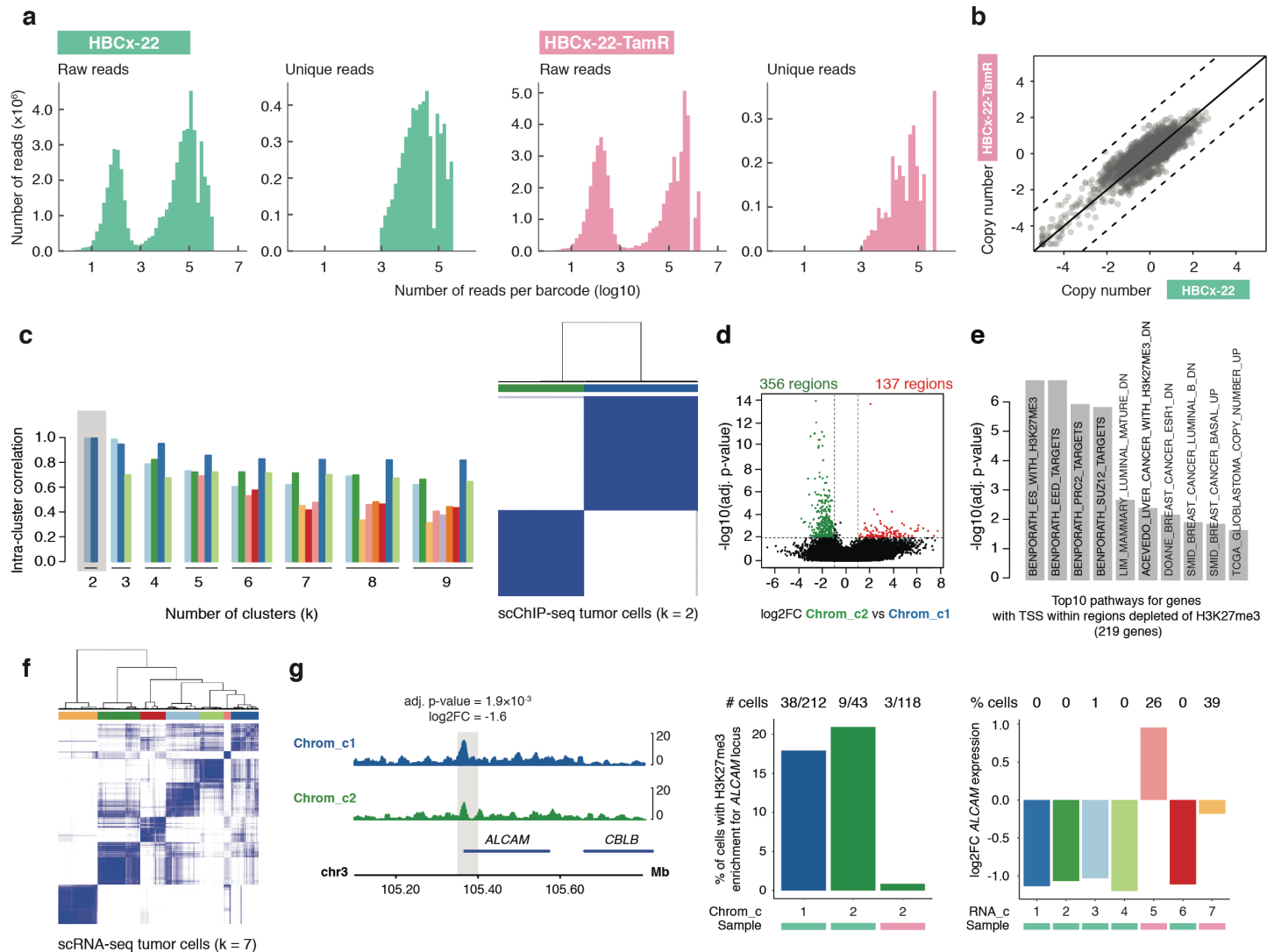
(a) Hierarchical clustering and corresponding heatmap of cell-to-cell Pearson's correlation scores. Cluster membership from consensus clustering, sample of origin (green for HBCx-95 [n = 617 single-cells] and pink for HBCx-95-CapaR [n = 926 single-cells]) and unique read count are indicated above the heatmap. (b) Consensus clustering scores for hierarchical clustering of stromal cells. Consensus score ranges from 0 (white: never clustered together) to 1 (dark blue: always clustered together). Cluster membership is color coded above the dendrogram. (c) t-SNE plots with dots colored according to the expression signal of each marker gene, representative of each cell population, in each cell.



Supplementary Figure 13

Clustering of single-cell ChIP-seq H3K27me3 profiles and scRNA-seq profiles of human tumor cells from the HBCx-95 model

(a) Left: Plot of copy number in 0.5 Mb non-overlapping regions in Capecitabine-resistant PDX (HBCx-95-CapaR) versus untreated PDX (HBCx-95), obtained from the input of the bulk ChIP-seq experiments. Regions with a deviation greater than $n = 2$ standard deviations (dashed black lines) compared to the mean (diagonal black line) were removed for subsequent scChIP-seq analysis. Middle and right: snapshots of loci affected by copy number variation for bulk DNA profiles of Capecitabine-resistant PDX and untreated PDX indicated in gray. (b) Left panel: hierarchical clustering and corresponding heatmap of cell-to-cell Pearson's correlation scores. Sample of origin (green for HBCx-95 [$n = 474$ single-cells] and pink for HBCx-95-CapaR [$n = 838$ single-cells]) and unique read count are indicated above the heatmap. Middle panel: t-SNE plots of scRNA-seq tumor cells, dots are colored according to the sample of origin and consensus clustering segmentation. Right panel: Consensus clustering scores for hierarchical clustering of scRNA-seq tumor cells. Consensus score ranges from 0 (white: never clustered together) to 1 (dark blue: always clustered together). Cluster membership is color coded beneath the dendrogram. (c) Consensus clustering analysis for scChIP-seq dataset. Left panel: mean of all pairwise correlation score between cluster's members is plotted for k clusters ranging from 2 to 10. At $k = 2$ clusters, the intra-cluster correlation is maximized. Right panel: hierarchical clustering and corresponding heatmap of cell-to-cell consensus clustering scores for scChIP-seq on tumor cells (HBCx-95 and HBCx-95-CapaR PDXs). Consensus scores ranges from 0 (white: never clustered together) to 1 (dark blue: always clustered together). Cluster membership is color coded above heatmap. (d) Barplot displaying the $-\log_{10}$ of adjusted P values from pathway analysis for regions with depletion of H3K27me3 in resistant cells. The gene sets are indicated on the barplot. (e) Left panels: Aggregated H3K27me3 chromatin profiles for *Chrom_c1* and *Chrom_c2* ($n = 457$ and 794 single-cells for *Chrom_c1* and *Chrom_c2* respectively) are shown for the loci identified in Figure 3f, as significantly differentially enriched and expressed (two-sided Wilcoxon signed-rank test). For each window indicated in gray, the $\log_2$ fold-change, the adjusted P value, the number and the proportion of cells with H3K27me3 enrichment within each cluster are indicated. Right panels: t-SNE representation of scRNA-seq datasets. Dots are colored according to expression signal in each cell.



Supplementary Figure 14

Clustering of single-cell ChIP-seq profiles of human tumor cells from the HBCx-22 model

(a) Histograms of the distribution of scChIP-seq raw and unique sequencing reads per cell in untreated HBCx-22 and Tamoxifen-resistant HBCx-22-TamR PDX. (b) Copy number in 0.5 Mb non-overlapping regions plotted for bulk DNA profiles of Tamoxifen-resistant PDX (HBCx-22-TamR) versus untreated PDX (HBCx-22). No aberrant variation in copy number was identified in this xenograft model between the resistant and sensitive tumors. The mean (diagonal black line) and deviations greater than $n = 2$ standard deviations compared to the mean (dashed black lines) are represented. (c) Consensus clustering analysis for scChIP-seq dataset. Left panel: mean of all pairwise correlation score between cluster's members is plotted for k clusters ranging from 2 to 10. At $k = 2$ clusters, the intra-cluster correlation is maximized. Right panel: hierarchical clustering and corresponding heatmap of cell-to-cell consensus clustering scores for scChIP-seq on tumor cells (HBCx-22 and HBCx-22-TamR PDXs). Consensus scores ranges from 0 (white: never clustered together) to 1 (dark blue: always clustered together). Cluster membership is color coded above heatmap. (d) Volcano plot representing adjusted P values (two-sided Wilcoxon signed-rank test) versus fold-changes for differential analysis comparing chromatin marks between *Chrom_c2* and *Chrom_c1* ($n = 356$ enriched and 137 depleted regions with adjusted P value < 0.01 and $|\log_2FC| > 1$). (e) Barplot displaying the $-\log_{10}$ of adjusted P values from pathway analysis for regions with depletion of H3K27me3 in cells from *Chrom_c2*. The gene sets are indicated on the barplot. (f) Hierarchical clustering and corresponding heatmap of cell-to-cell consensus clustering score for scRNA-seq tumor cells (HBCx-22 and HBCx-22-TamR PDXs). Consensus score ranges from 0 (white: never clustered together) to 1 (dark blue: always clustered together). Cluster membership is color coded above the heatmap. (g) Left panel: aggregated H3K27me3 chromatin profiles for *Chrom_c1* and *Chrom_c2* are shown for the *ALCAM* locus ($n = 212$ and 161 single-cells for *Chrom_c1* and *Chrom_c2* respectively). Differentially bound region identified by two-sided Wilcoxon signed-rank test is highlighted in gray with the corresponding adjusted P value and $\log_2$ fold-change. Middle panel: barplot displaying the proportion of cells with H3K27me3 enrichment in each cluster. The corresponding number of cells is indicated above the barplot. For each cluster, the origin of cells (green for HBCx-22 and pink for HBCx-22-TamR) is indicated below. Right panel: barplot displaying the average $\log_2$ fold-change for *ALCAM* expression level for cells in each cluster versus all remaining cells. The percentage of cells, within each cluster, with detectable *ALCAM* expression is indicated above the barplot. For each cluster, the origin of cells (green for HBCx-22 and pink for HBCx-22-TamR) is indicated below.