



HAL
open science

Interactive analysis of single-cell epigenomic landscapes with ChromSCape

Pacôme Prompsy, Pia Kirchmeier, Justine Marsolier, Marc Deloger, Nicolas
Servant, Céline Vallot

► **To cite this version:**

Pacôme Prompsy, Pia Kirchmeier, Justine Marsolier, Marc Deloger, Nicolas Servant, et al.. Interactive analysis of single-cell epigenomic landscapes with ChromSCape. *Nature Communications*, 2020, 11, 10.1038/s41467-020-19542-x . hal-03075604

HAL Id: hal-03075604

<https://hal.science/hal-03075604>

Submitted on 16 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Interactive analysis of single-cell epigenomic landscapes with**
2 **ChromSCape**

3

4 Pacôme Prompsy^{1,2}, Pia Kirchmeier^{1,2}, Justine Marsolier^{1,2}, Marc Deloger³, Nicolas Servant³,
5 Céline Vallot^{1,2}

6

7

8 ¹CNRS UMR3244, Institut Curie, PSL Research University, 26 rue d'Ulm, 75005 Paris, France,

9 ²Translational Research Department, Institut Curie, PSL Research University, 26 rue d'Ulm,

10 75005 Paris, France

11 ³INSERM U900, Institut Curie, PSL Research University, Mines ParisTech, 26 rue d'Ulm,

12 75005 Paris, France

13

14

15

16

17

18 Correspondence to: pacome.prompsy@curie.fr, celine.vallot@curie.fr

19

20 **Abstract**

21 Chromatin modifications orchestrate the dynamic regulation of gene expression during
22 development and in disease. Bulk approaches have characterized the wide repertoire of histone
23 modifications across cell types, detailing their role in shaping cell identity. However, these
24 population-based methods do not capture cell-to-cell heterogeneity of chromatin landscapes,
25 limiting our appreciation of the role of chromatin in dynamic biological processes. Recent
26 technological developments enable the mapping of histone marks at single-cell resolution,
27 opening up perspectives to characterize the heterogeneity of chromatin marks in complex
28 biological systems over time. Yet, existing tools used to analyze bulk histone modifications
29 profiles are not fit for the low coverage and sparsity of single-cell epigenomic datasets. Here,
30 we present ChromSCape, a user-friendly interactive Shiny/R application that processes single-
31 cell epigenomic data to assist the biological interpretation of chromatin landscapes within cell
32 populations. ChromSCape analyses the distribution of repressive and active histone
33 modifications as well as chromatin accessibility landscapes from single-cell datasets. Using
34 ChromSCape, we deconvolve chromatin landscapes within the tumor micro-environment,
35 identifying distinct H3K27me3 landscapes associated to cell identity and breast tumor subtype.

36 **Introduction**

37 Histone modifications are key regulators of gene expression, driving chromatin folding and
38 genes accessibility to transcription machineries. The recent development of single-cell methods
39 to study epigenomes now enables the appreciation of the heterogeneity of chromatin
40 modifications within a population. These experimental methods assess the distribution of
41 histone marks at single-cell resolution by coupling next generation sequencing to high-
42 throughput microfluidics DNA barcoding (scChIP-seq)^{1,2} or *in situ* reactions (scChIL-seq³,
43 scChIC-seq⁴, scCUT&Tag⁵). In contrast to scATAC-seq approaches which identify open
44 regions of the chromatin⁶⁻⁸, these methods can capture various chromatin states, enriched in
45 repressive or active histone marks (H3K27me3 or H3K4me3 for example). Using these
46 approaches, we can study the heterogeneity of epigenomes within complex biological samples,
47 such as tumors¹, and start appreciating the role of epigenomic diversity and the dynamics of
48 chromatin in disease and development.

49 Existing tools used to analyze bulk ChIP-seq experiments are not fit for the low coverage and
50 sparsity of these single-cell histone modifications datasets, which is due to the inherent low
51 number of copy of DNA molecules per cell - maximum two for a diploid genome. Several
52 computational methods for the analysis of scATAC-seq have been developed to deal with the
53 specificities of single-cell DNA-based datasets. They were recently benchmarked⁹, with
54 SnapATAC¹⁰, CisTopic¹¹ and Cusanovich2018¹² being the top-three performing methods.
55 These tools, initially dedicated to scATAC-seq and without graphic interface, require some
56 scripting skills. Biologists with limited computational training can manipulate and analyze
57 scRNA-seq and scATAC-seq datasets using applications such as 'scOrange'¹³ and 'SCRAT'¹⁴.
58 With ChromSCape (Figure 1), we propose a user-friendly, step-by-step and customizable
59 Shiny/R application to analyze all types of sparse single-cell epigenomic datasets. The user can
60 interactively identify subpopulations with common epigenomes within heterogeneous samples,

61 find differentially enriched regions between subpopulations and interpret epigenomes by
62 linking regions to associated genes and pathways. The pipeline starts from aligned sequences
63 or count tables, and is designed for high-throughput single-cell datasets with samples containing
64 as low as 100 cells with a minimum of 1,000 reads per cell up to 25,000 cells on a standard
65 laptop. ChromSCape accepts multiple samples to allow comparisons of cell populations
66 between and within samples. ChromSCape can determine cell identities from single-cell histone
67 modification profiles, whatever the technology, as well as scATAC-seq datasets. We showcase
68 the use of ChromSCape by deconvolving chromatin landscapes within the tumor micro-
69 environment; we identify distinct H3K27me3 landscapes associated to cell identity and breast
70 tumor subtype.

71

72 **Results**

73 **ChromSCape identifies cell identities from scChIP-seq data**

74 To test the efficiency of ChromSCape (Figure 1) in identifying cell sub-populations based on
75 their epigenome (H3K27me3), we generated an *in-silico* dataset with known ground truth,
76 mixing 4 different human cell types: Jurkat B cells, Ramos T cells, MDA-MB-468 breast cancer
77 cells and HBCx-22 tumor cells derived from a luminal breast tumor PDX model¹. Interestingly,
78 Jurkat and Ramos cells were processed within the same microfluidics experiment, preventing
79 the existence of any batch effect between them (see Grosselin et al., ¹). We compared
80 ChromSCape to methods specifically designed for single-cell epigenomic datasets (scATAC-
81 seq) for their ability to identify cell identities. Based on a recent scATAC-seq benchmark⁹, we
82 selected the top-performing methods, namely *Cusanovich2018*¹², SnapATAC¹⁰ and CisTopic¹¹.
83 We also benchmarked EpiScanpy³², a recent analysis pipeline for various single-cell
84 epigenomic data (scATAC-seq, scDNA methylation, ...) developed in Python. We applied

85 hierarchical clustering on the reduced feature space obtained by each method and used an ARI
86 metric to evaluate their ability to identify cell phenotypes. ChromSCape with default parameters
87 manages to separate almost perfectly the 4 cell types, with an ARI of 0.998 (Figure 2a), as
88 *Cusanovich2018* and CisTopic (both an ARI of 0.996, Figure 2b), followed closely by
89 EpiScanpy (ARI of 0.940, Figure 2b). ChromSCape, EpiScanpy and SnapATAC were all run
90 on 50kbp bins, but SnapATAC had noisier clusters and a slightly poorer ARI (0.822).

91 We also compared the agility of ChromSCape to manipulate and interpret scChIP-seq datasets
92 to two applications with graphic interface, using the same reference dataset (Figure 2c), with
93 either default settings or optimizing input and settings. scOrange is a stand-alone platform
94 allowing researchers to create workflows to analyze single-cell datasets, offering a wide variety
95 of analytical modules. While for scRNA-seq many workflows have been developed and are
96 ready-to-use, in the case of epigenomic datasets, users need to have prior computational
97 knowledge to organize a proper workflow. We managed to group cells according to sample of
98 origin only with an optimized workflow (Figure 2c, default vs optimized). SCRAT is a Shiny/R
99 package presented as a user interface to analyze single-cell epigenomic data. The default option
100 for SCRAT is to count reads within ‘ENCODE Clusters’ corresponding to co-regulatory open
101 chromatin regions obtained from DNase-seq datasets, not adapted for the analysis of repressive
102 histone marks like H3K27me3. In order to use SCRAT for our reference scChIP-seq datasets,
103 we had to pre-compute counts on pre-defined peaks called on the ‘pseudo-bulk’ (see Methods
104 and ‘Optimized’ panel), limiting the usability of SCRAT. In addition, in contrast to
105 ChromSCape, both applications do not propose functionalities to associate genomic regions to
106 gene annotation, limiting the biological interpretation of the results obtained with differential
107 analysis.

108 Like single-cell transcriptomics approaches, single-cell epigenomic technologies can be
109 influenced by various batch effects, e.g. library preparation, batch of hydrogel beads or efficacy

110 of immuno-precipitation or cleavage. To overcome this, we have implemented in ChromSCape
111 a module for batch correction based on the fastMNN method¹⁸. To test this functionality, we
112 used three datasets: two from our previous study (HBCx-95 and HBCx95-CapaR, collected
113 with batch1 beads) and a new dataset, HBCx95 – batch 2, which is a biological replicate of
114 HBCx-95, i.e a PDX tumor from the same PDX model but a different mouse, processed with a
115 new batch of beads. Using ChromSCape, we analyzed together the three H3K27me3 tumor
116 samples. As shown in Supplementary Figure 2, a strong batch effect separates the two biological
117 replicates before batch correction (left panel). After applying batch correction (right panel),
118 cells from the two biological replicates of untreated HBCx-95 tumors successfully mix, but not
119 with cells from the resistant tumor, suggesting that the module corrects for batch effect without
120 overcorrecting biological differences.

121 We also evaluated the usability of ChromSCape for other types of single-cell histone
122 modification data obtained by other technologies than scChIP-seq. We analyzed two public
123 datasets of scCUT&Tag and scChIC-seq targeting H3K27me3 and H3K4me3 marks
124 respectively. ChromSCape facilitates the analysis of such public dataset as the user can directly
125 upload the GEO single-cell BED files into the application. We recommend here for H3K27me3
126 mark – accumulating in broad peaks - to aggregate the signal into 50kbp bins, and for H3K4me3
127 mark – accumulating in sharp peaks - to count within 5kbp bins or around gene TSS (+/-
128 2500bp). As shown in Supplementary Figure 1a, for the scCUT&Tag dataset the two K562
129 replicates showed no batch effect and were clustered together separately from H1 cells (ARI =
130 0.976). For the scChIC-seq dataset (Supplementary Figure 1b), 7 clusters are clearly observable
131 on the UMAP representation, as was found by the authors in their study⁴.

132

133 **ChromSCape classifies cells from scATAC-seq data**

134 In order to assess the capacity of ChromSCape to analyze all types of single-cell epigenomic
135 data, we re-analysed a scATAC-seq dataset (GSE99172) containing 8 cell lines and 4 patient-
136 derived cells. This dataset was partly produced and analyzed in a study by chromVar, a
137 dedicated scATAC-seq analytical tool³¹; we used the same color code as in this study. This
138 dataset contains various biological samples as well as technical replicates for two cell lines,
139 K562 & GM12878, for which there are 6 and 4 technical replicates respectively. Due to a
140 relatively low number of cells per sample (n=96 per sample), we set the read count threshold to
141 1,000 for cells to be included in the analysis. We measured the ability of ChromSCape to
142 classify cells according to cell type of origin using assignment scores for each sample X and
143 each cluster Y (*number of cells from sample X assigned to cluster Y / total number of cells of*
144 *sample X*).

145 In the unsupervised analysis, we identified the optimal number of clusters to be $k = 5$ according
146 to the relative change in area under the CDF curve (Figure 3a and ~~data not shown~~
147 Supplementary Figure 3a). Cells from different technical replicates of K562 and GM12878 all
148 grouped together in clusters 5 and 2 with assignment scores of 99.7% and 98.3% respectively
149 (Figure 3b). Analyzing all samples together, ChromSCape could robustly identify - i.e as stable
150 separated clusters by consensus clustering - TF1, K562 and GM12878 cells, affecting in
151 average 99.4% of cells to correct cluster. Cluster C1 grouped together AML, Mono and LMPP
152 samples with HL60 cell line, which is also originally derived from leukocytes of a patient with
153 an AML cancer.

154 In order to get more insight into cell identities within cluster C1, we re-analyzed cells from C1
155 with ChromSCape. In contrast to the first round of analysis, ChromSCape was able to
156 distinguish cell identities within samples, and detect individual clusters of cells, with high

157 assignment scores for the two normal samples (LMPP & Monocytes) and HL-60 (average
158 assignment score 98.3%, Figure 3c & d). Additionally, AML blasts from patient SU070 show
159 a greater monocyte signature than patient SU353 (Figure 3d, p-value = 0.0025, Fisher's exact
160 test, respectively 85.0 % and 25.7 % of SU070 and SU353 blasts cells cluster with monocytes),
161 as previously described for these cells in ²⁹. ChromSCape identifies distinct populations within
162 the normal immune cell environment based on their chromatin accessibility. Within AML
163 patient samples, ChromSCape matches each cancer cell to the closest resembling healthy
164 population.

165

166 **ChromSCape deconvolves epigenomes of the tumor micro-environment**

167 To further showcase the use of ChromSCape, we interrogated the heterogeneity of chromatin
168 states within the tumor micro-environment of two breast tumor subtypes: luminal and triple-
169 negative (TNBC) breast tumors. Tumor micro-environment is a key player in tumor evolution
170 processes, and can vary between tumor types and with response to cancer therapy. Here our
171 goal was to compare H3K27me3 landscapes of cells from the tumor micro-environment of
172 luminal and TNBC subtypes, resistant or not to cancer treatment. The HBCx-22 and HBCx-22-
173 TamR datasets correspond to mouse cells from a pair of luminal ER⁺ breast PDXs¹: HBCx-22,
174 responsive to Tamoxifen and HBCx-22-TamR, resistant to Tamoxifen. The HBCx-95 and
175 HBCx-95-CapaR correspond to a triple-negative breast cancer (TNBC) tumor model of
176 acquired resistance to chemotherapy¹. We analyzed together these four H3K27me3 mouse
177 scChIP-seq datasets, two of which had not been analyzed in our previous study¹. Using
178 ChromSCape, we propose a comprehensive view of cell populations based on their chromatin
179 profiles, and show the identification of tumor-type and treatment-specific cell populations and
180 respective chromatin features. All plots in Figure 4 were automatically generated by the

181 application and are downloadable from the interface. In the quality filtering step, a threshold of
182 2,000 reads per cell was set due to a relatively high initial number of cells ($n = 5,516$ cells).
183 After the dimensionality reduction step (Figure 4a & b), we applied our consensus clustering
184 approach on the filtered dataset with $k = 2$ to $k = 10$ clusters. We chose to partition the data into
185 $k = 4$ clusters based on the knee method, as a plateau in the relative change in area under the
186 CDF curve was observed between $k = 4$ and $k = 5$ clusters (Figure 4c-d & Supplementary Figure
187 34a-b). Consensus score matrix in Figure 4d shows that most of the cells were stably assigned
188 to four chromatin-based populations (mean consensus score for elected clusters of 0.91 is
189 significantly higher than mean consensus score for other clusters, 0.17, p -value = $2.2e-16$, two-
190 sided Student's t -test). Assignment of cells to cluster C2 and C4 is significantly less stable than
191 C1 and C3 (p -value < $2.2e-16$, Student's two-sided t -test, mean consensus scores are
192 respectively 0.84 and 0.90 for C1-C3 and 0.70 and 0.71 for C2-C4, see Supplementary Figure
193 3a), suggesting that cells from C2 and C4 might share H3K27me3 features, whereas cells from
194 C1 and C3 have distinct H3K27me3 landscapes. Clusters C1, C2 and C4 contain cells from all
195 four samples, with a significantly higher proportion of HBCx-22-TamR for C1 (p -value = $3e$ -
196 05, Pearson's Chi-squared test) (Figure 4e). On the other hand, cluster C3 is almost exclusively
197 composed of cells from model HBCx-95 (Figure 4c&e), revealing a stromal cell population
198 specific to the triple negative breast cancer model (HBCx-95).

199 To further identify the specific features of each chromatin-based population, we proceeded to
200 peak calling, differential analysis and gene set enrichment analysis using default parameters
201 (see Methods). As H3K27me3 is a repressive histone mark, we focused our analysis on loci
202 depleted in H3K27me3, where transcription of genes can occur, in cells from each cluster versus
203 all other cells. The differential analysis identified respectively 189, 210, 83 and 9 significantly
204 depleted regions for clusters C1 to C4 (Figure 4g, $\log_{2}FC < 1$, adjusted p -value < 0.01). We
205 found loci devoid of H3K27me3 specific to cluster C2, enriched for genes involved in apical

206 junction such as *Bcar1* (Figure 4f) and *Ptk2*, which are characteristic of genes expressed in
207 fibroblasts. We found a depletion of H3K27me3 specific to cluster C3 over the genes *Nrros*
208 (Figure 4f) and *Il10ra*, two genes characteristic of immune expression programs. Depletion of
209 H3K27me3 over the transcription start site of *Rap1gap2*, a gene expressed in endothelial cells,
210 was a key feature of cluster C4 (Figure 4f). For cluster C1 and C2, we found a depletion of
211 H3K27me3 over *Eln*, a gene expressed in fibroblasts.

212 Gene set enrichment analysis for genes located in regions depleted of H3K27me3 enrichment
213 only revealed very few enriched gene lists, mostly for cluster C2 (q-value < 0.1, Figure 4h,
214 multiple gene sets related to stem and cancer cells) and one list for C1
215 (“LPS_VS_CONTROL_MONOCYTE_UP”). Linking H3K27me3 enrichment to transcription
216 is indeed indirect, we envisage such enrichment analysis more appropriate for H3K4me3
217 scChIP-seq in which enriched regions are directly associated to gene transcription.

218 Overall, these results are consistent with our previous analysis of HBCx-95 scRNA-seq datasets
219 where subpopulations were differentially expressing markers of fibroblasts, endothelial and
220 macrophage cells¹. This new analysis comprising the HBCx-22 dataset allowed us to identify
221 the H3K27me3 signature of potential endothelial cells (cluster C4). These cells are present in
222 each model, but might not have been previously detected in the previous scChIP-seq analysis
223 due to low cell representation. In addition, the H3K27me3 signature of potential immune cells
224 is restricted to cells from the TNBC model (cluster C3), suggesting that these immune cells are
225 absent from the luminal tumor.

226

227 **Discussion**

228 ChromSCape is a Shiny/R application designed for both biologists and bioinformaticians to
229 analyze complex chromatin profiling datasets such as scChIP-seq datasets. The comprehensive

230 application is quick to take over plus the direct visualization of cells clusters combined to
231 configurable parameters and incremental saving of intermediary R objects eases bench-marking
232 of parameters. We show that ChromSCape performs as well or better than state of the art single-
233 cell epigenomic analytic tools to identify cell identities from an *in-silico* mix of H3K27me3
234 scChIP-seq datasets. It also manages to identify sub-populations within a complex scATAC-
235 seq benchmarking dataset, showing its wide range of application for epigenomic analysis. In
236 addition, using ChromSCape to study the epigenome of mouse stromal cells in breast tumors,
237 we can identify the various epigenomes within the tumor micro-environment. Overall, we see
238 ChromSCape as a useful tool to probe heterogeneity and dynamics of chromatin profiles in
239 various biological settings, not only in cancer development but also in cell development and
240 cellular differentiation.

241

242 **Methods**

243 **Implementation**

244 ChromSCape is an R package developed in Shiny/R. It uses various Shiny related packages
245 (shinyjs, shinydashboard, shinyDirectoryInput) for the user interface. The application takes
246 advantage of public R libraries for data vizualisation (RcolorBrewer, colorRamps, Rtsne, umap,
247 colourpicker, kableExtra, knitr, viridis, ggplot2, gplots, png, grid, gridExtra, DT) as well as for
248 data manipulation (Matrix, dplyr, tidyr, stringr, irlba, rlist, qualV, stringdistr). ChromSCape
249 uses Bioconductor packages (i) for the manipulation of single-cell data with
250 SingleCellExperiment, scater¹⁵, scran¹⁶, (ii) for the manipulation of genomic regions with
251 IRanges and GenomicRanges¹⁷, (iii) for the manipulation of genomic files with Rsamtools and
252 BiocParallel, (iv) for the correction of batch effects with batchelor¹⁸ and (v) to determine the
253 optimal number of clusters with ConsensusClusterPlus¹⁹. In addition, ChromSCape makes use
254 of custom R functions which serve for both manipulation and visualization of datasets. Brief
255 command lines enable users without any bioinformatics skills to install all R dependencies and
256 run the application in a web browser.

257

258 **Demonstration application**

259 A demonstration of ChromSCape is freely available at
260 <https://vallotlab.shinyapps.io/ChromSCape/>.

261

262 **Input datasets, quality control and pre-processing**

263 Input files for ChromSCape are either one or multiple count matrices with genomic regions in
264 rows and cells in columns or single-cell BAM or BED files. In this case, a directory containing

265 single-cell BAM or BED files must be specified and the count matrix created by aggregating
266 the signal into successive genomic bins, peaks (BED file must be provided by user) or into
267 regions around genes Transcription Start Sites (TSS). For H3K27me3 scChIP-seq datasets, with
268 a distribution in broad peaks, we recommend using bins of 50kbp, while for H3K4me3 scChIP-
269 seq or scATAC-seq datasets we recommend using smaller bins (e.g 5kb), known peaks or
270 regions around TSS. The ‘condition’ or ‘label’ of each cell is then heuristically determined
271 using file names and the number of conditions specified by the user. Guidelines and link
272 towards datasets are given in the user guide
273 (https://vallotlab.github.io/ChromSCape/ChromSCape_guide.html).

274 In order to efficiently remove outlier cells from the analysis, e.g. cells with excessively high or
275 low coverage, the user sets a threshold on a minimum read count per cell and the upper
276 percentile of cells to remove. The latter could correspond to doublets, e.g. two cells in one
277 droplet, while lowly covered cells are not informative enough or may correspond to barcodes
278 ligated to contaminant DNA or library artifacts. Regions not supported by a minimum user-
279 defined percentage of cells that have a coverage greater than 1,000 reads are filtered out.
280 Defaults parameters were chosen based on the analysis of multiple scChIP-seq datasets from
281 our previous study¹: a minimum coverage of 1,600 unique reads per cell, filtering out the cells
282 with the top 5% coverage and keeping regions detected in at least 1 % of cells. Post quality
283 control filtering, the matrices are normalized by total read count and region size. At this step,
284 the user can provide a list of genomic regions, in BED format, to exclude from the subsequent
285 analysis, in cases of known copy number variation regions between cells for example.

286 To reduce the dimensions of the normalized matrix for further analysis, principle component
287 analysis (PCA) is applied to the matrix, with centering, and the 50 first PCs are kept for further
288 analysis. The user can visualize scChIP-seq data after quality control in the PCs dimensional
289 space. The t-distributed stochastic neighbor embedding (t-SNE) algorithm²⁰ and UMAP²¹ is

290 applied on the PCA to visualize the data in two dimensions. The PCA and t-SNE plots are a
291 convenient way to check if cells form clusters in a way that was expected before any clustering
292 method is applied. For instance, the user should verify whether the QC filtering steps and
293 normalization procedure were efficient by checking the distribution of cells in PC1 and PC2
294 space. Cells should group independently of normalized coverage. In our hands, for our scChIP-
295 seq H3K27me3 datasets, minimum coverage of 1,600 unique reads per cell was required to
296 separate cells independently of coverage post normalization¹. A batch correction option using
297 mutual nearest neighbors ‘FastMNN’ function from ‘batchelor’ package¹⁸ is implemented to
298 remove any known batch effect in the reduced feature space.

299

300 **Hierarchical clustering, filtering and consensus clustering**

301 Using the first 50 first PCs of computed PCA as input, hierarchical clustering is performed,
302 taking 1-Pearson’s correlation score as distance metric. To improve the stability of our
303 clustering approaches and to remove from the analysis isolated cells that do not belong to any
304 subgroup, cells displaying a Pearson’s pairwise correlation score below a threshold t with at
305 least p % of cells are filtered out (p is set at 1 % by default). The correlation threshold t is
306 calculated as a user-defined percentile of Pearson’s pairwise correlation scores for a randomized
307 dataset (percentile is recommended to be set as the 99th percentile). Correlation heatmaps
308 before and after correlation filtering and the number of remaining cells are displayed to inform
309 users on the filtering process.

310 ChromSCape uses Bioconductor ConsensusClusterPlus package¹⁹ to determine what is the
311 appropriate k -partition of the filtered dataset into k clusters. To do so, it evaluates the stability
312 of the clusters and computes item consensus score for each cell for each possible partition from
313 $k=2$ to 10. For each k , consensus partitions of the dataset are done on the basis of 1,000
314 resampling iterations (80% of cells sampled at each iteration) of hierarchical clustering, with

315 Pearson's dissimilarity as the distance metric and Ward's method for linkage analysis. The
316 optimal number of clusters is then chosen by the user; one option is to maximize intra-cluster
317 correlation scores based on the graphics displayed on the 'Consensus Clustering' tab after
318 processing. Clustering memberships can be visualized in two dimensions with the t-SNE or
319 UMAP plot.

320

321 **Peak calling for genomic region annotation**

322 This step of the analysis is optional, but recommended in order to refine the peak annotation
323 prior to enrichment analysis. To be able to run this module, some additional command line tools
324 are required such as SAMtools²², BEDTools²³ and MACS2²⁴. The user needs to input BAM
325 files for the samples (one separate BAM file per sample), with each read being labeled with the
326 barcode ID. ChromSCape merges all files and splits them again according to the previously
327 determined clusters of cells (one separate BAM file per cluster). Customizable significance
328 threshold for peak detection and merging distance for peaks (defaults to p-value=0.05 and peak
329 merge distance to 5,000) allows to identify peaks in close proximity (<1,000bp) to a gene
330 transcription start site (TSS); these genes will be later used as input for the enrichment analysis.
331 For the annotation, ChromSCape uses the reference human transcriptome Gencode_hg38_v26,
332 limited to protein coding, antisense and lncRNA genes.

333

334 **Differential Analysis and pathway enrichment analysis**

335 To identify differentially enriched regions across single-cells for a given cluster, ChromSCape
336 can perform (i) a non-parametric two-sided Wilcoxon rank sum test comparing normalized
337 counts from individual cells from one cluster versus all other cells, or cluster of choice, or (ii)
338 a parametric test comparing raw counts from individual cells, using edgeR²⁵, based on the

339 assumption that the data follows a negative-binomial distribution. We test for the null
340 hypothesis that the distribution of normalized counts from the two compared groups have the
341 same median, with a confidence interval 0.95. The calculated p-values are then corrected by the
342 Benjamini-Hocheberg procedure²⁶. The user can set a log2 fold-change threshold and corrected
343 p-value threshold for regions to be considered as significantly differentially enriched (default
344 settings are a p-value and log2 fold-change thresholds respectively of 0.01 and 1). If users have
345 specified batches, the differential analysis is done using the ‘pairwiseWilcox’ function from the
346 scran package¹⁶, setting the batch of origin as a ‘blocking level’ for each cell.
347 For the top 100 most significant differential regions, single-cell H3K27me3 enrichment levels
348 can be visualized overlaying H3K27me3 counts for each cell at selected genes onto a t-SNE
349 plot. Using the refined annotation of peaks done in previous step, the final step is to look for
350 enriched gene sets of the MSigDB v5 database²⁷ within differentially enriched regions (either
351 enriched or depleted regions in the studied histone mark). We apply hypergeometric tests to
352 identify gene sets from the MSigDB v5 database over-represented within differentially enriched
353 regions, correcting for multiple testing with the Benjamini-Hochberg procedure. Users can then
354 visualize most significantly enriched or depleted gene sets corresponding to the epigenetic
355 signatures of each cluster and download gene sets enrichment tables.

356

357 **Datasets**

358 H3K27me3 scChIP-seq human *in-silico* mix of 4 cell types: The samples correspond to n=326
359 human tumor cells from untreated PDX (HBCx-22), n=201 human T cells (Jurkat) and n=306
360 B cells (Ramos) taken from ¹ and n=454 cells from the MDA-MB-468 triple-negative breast
361 cancer cell line (HBCx-22, Jurkat and Ramos data are from [GSE117309](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117309)
362 [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117309>]), MDA-MB-468 is
363 available at [GSE152502](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152502) [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152502>]).

364 H3K4me3 scChIC-seq human white blood cells dataset⁴: n=285 white blood cells from a human
365 male donor were downloaded as gzipped single-cell BED files from GSE105012
366 [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE105012>], inputted directly into
367 ChromSCape and aggregated into 50kbp bins (default).

368 H3K27me3 scCUT&Tag human H1 and K562 cells⁵: A replicate of K562 cell line comprising
369 of n=908 cells from GSE124680
370 [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124680>], another replicate of
371 n=479 K562 cells and n=486 H1 cells from GSE124690
372 [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124690>] were downloaded as
373 gzipped single-cell BED files, inputted directly into ChromSCape and aggregated around gene
374 TSS (+/- 2500bp).

375 H3K27me3 scChIP-seq human datasets: The samples correspond to human cells from patient-
376 derived xenograft (PDX) originating from two different human donors¹. For this study, we
377 added a new scChIP-seq dataset, corresponding to a biological replicate of HBCx-95
378 (GSE152502 [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152502>]), processed
379 with a novel batch of hydrogel beads.

380 scATAC-seq datasets: The scATAC-seq dataset is composed of two cell types derived from
381 two acute myeloid leukaemia (AML) (patients SU070 and SU353 blastocytes (blast) and
382 leukemic stem cells (LSC) from ²⁹) as well as multiple cell lines : GM12878 (4 replicates),
383 TF1, BJ, H1, HL60, K562 (3 replicates) from ³⁰, K562 (3 replicates) from ³¹; monocytes (Mono)
384 and lymphoid primed multipotent progenitor (LMPP) from ²⁹. The count matrix of reads in
385 peaks was downloaded from GEO accession number GSE99172
386 [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE99172>], split into distinct matrices
387 for each sample and formatted to be accepted as input by ChromSCape.

388 H3K27me3 scChIP-seq mouse datasets: The samples correspond to mouse cells from patient-
389 derived xenograft (PDX) originating from two different human donors¹. Raw FASTQ reads
390 were processed using the latest version of our scChIP-seq data engineering pipeline (see above)
391 to produce 50kbp binned count matrices given as input to ChromSCape (matrices available at
392 [https://figshare.com/projects/Single-Cell_ChIP-
393 seq_of_Mouse_Stromal_Cells_in_PDX_tumour_models_of_resistance/66419](https://figshare.com/projects/Single-Cell_ChIP-seq_of_Mouse_Stromal_Cells_in_PDX_tumour_models_of_resistance/66419)).

394

395 **Cell line**

396 MDA-MB-468 cells, bought at ATCC (HTB-132™), were cultured in DMEM 1640 (Gibco-
397 BRL) and supplemented with 10% heat-inactivated fetal calf serum. Cell numbers, as judged
398 by Trypan Blue exclusion test, were determined by counting cells using a Countess automated
399 cell counter (Invitrogen). Cells were cultured at 37 °C in a humidified 5% CO₂ atmosphere.
400 The cell line was mycoplasma negative. The MDA-MB-468 cells were trypsinized (Trypsin,
401 Gibco-BRL). Prior to single-cell ChIP-seq, cells were then re-suspended in PBS/0.04% BSA
402 (ThermoFisher Scientific, # AM2616).

403

404 **Patient-derived xenograft (PDX)**

405 Female Swiss nude mice were purchased from Charles River Laboratories and were maintained
406 under specific pathogen-free conditions. Their care and housing were in accordance with
407 institutional guidelines and the rules of the French Ethics Committee (project authorization no.
408 02163.02). A PDX from a residual triple-negative breast cancer post neo-adjuvant
409 chemotherapy (HBCx-95) was previously established at Institut Curie with informed consent
410 from the patient¹. Prior to single-cell ChIP-seq, PDX were digested at 37°C for 2h with a
411 cocktail of Collagenase I (Roche, # 11088793001) and Hyaluronidase (Sigma, # H3506). Cells
412 were then individualized at 37°C using a cocktail of 0.25% trypsin/Versen (ThermoFisher

413 Scientific, #15040-033), Dispase II (Sigma, #D4693) and Dnase I (Roche, # 11284932001).
414 Red Blood Cell lysis buffer (ThermoFisher Scientific, # 00-4333-57) was then added to degrade
415 red blood cells. In order to increase the viability of the cell suspension, dead cells were removed
416 using the Dead Cell Removal kit (Miltenyi Biotec). Cells were re-suspended in PBS/0.04%
417 BSA (ThermoFisher Scientific, # AM2616).

418

419 **Single-cell ChIP-seq**

420 The protocol for scChIP-seq was rigorously the same as in Grosselin et al.,¹ and can be resumed
421 by the main following steps. Cells were first compartmentalized into droplets containing Mnase
422 in a microfluidics chip, then fused with barcoded hydrogel beads. After fusion of cell-
423 containing droplets and bead-containing droplets, Fast-link DNA ligase [Lucigen, # LK0750H]
424 was used to ligate segmented DNA to barcodes. Droplets were pooled and used for chromatin
425 immuno-precipitation against anti-H3K27me3 antibody ([Cell Signaling Technology, # 9733]).
426 After treatment with RNase A (ThermoFisher Scientific, #EN0531) and Proteinase K
427 (ThermoFisher Scientific, # EO0491), barcoded-nucleosomes were then amplified by in-vitro
428 transcription using the T7 MegaScript kit (ThermoFisher Scientific, # AM1334) and reverse-
429 transcribed. After RNA digestion, DNA was amplified by PCR. The final product was size-
430 selected by gel electrophoresis. Single-cell ChIP-seq libraries were finally sequenced on an
431 Illumina NextSeq 500 MidOutput 150 cycles.

432

433 **Demultiplexing and alignment of H3K27me3 scChIP-seq datasets**

434 Raw FASTQ reads were processed using the latest version of our scChIP-seq data engineering
435 pipeline that allowed a more precise removal of PCR and RT duplicates (code available at
436 https://github.com/vallotlab/scChIPseq_DataEngineering) to produce 50kbp binned count
437 matrices given as input to ChromSCape (matrices available at

438 https://figshare.com/projects/Single-Cell_ChIP-
439 [seq_of_Mouse_Stromal_Cells_in_PDX_tumour_models_of_resistance/66419](https://figshare.com/projects/Single-Cell_ChIP-)). Rapidly, the
440 first 56bp of the Read2 were separated into three indexes and aligned using bowtie2 separately
441 against reference of three pools of 96 16-bp long indexes. Reads containing all three
442 recognizable indexes (a full cell-barcode) were kept, the genomic part of Read2 and Read1 were
443 aligned in paired-end mode using STAR v2.7.0. For each barcode, aligned reads were
444 deduplicated by removing successively: (i) PCR duplicates, identified if #Read1 + #Read 2
445 mapped at the same position, (ii) RT duplicates, identified if #Read 1 mapped at the same
446 position and (iii) window duplicates: all the reads falling in the same 50 bp window were stacked
447 into one as reads possibly originating from the same nucleosome. Reads were binned in non-
448 overlapping 50 kb bins spanning the genome to generate a n x m coverage matrix with n
449 barcodes and m genomic bins used in downstream analysis.

450

451 **Benchmark of tools for scChIP-seq data analysis**

452 Three methods dedicated to the analysis of scATAC-seq with the best performance according
453 to Chen et al., 2019⁹, were tested on a mixture of H3K27me3 scChIP-seq datasets (see Datasets
454 below), namely ‘SnapATAC’, ‘CisTopic’ and ‘Cusanovich2018’. The scripts were taken from
455 the GitHub repository of the benchmark paper ([https://github.com/pinellolab/scATAC-](https://github.com/pinellolab/scATAC-benchmarking)
456 [benchmarking](https://github.com/pinellolab/scATAC-benchmarking)). For ‘CisTopic’ and ‘Cusanovich2018’, peaks were called using MACS2 with
457 options ‘--nomodel --extsize 300 --keep-dup all --broad’. Peaks closer to 5000bp were merged
458 together using BEDTools. For ‘SnapATAC’, 50kbp bins were counted from BAM files using
459 ‘SnapTools’. In addition, we also tested a recent method for single-cell epigenomic analysis,
460 ‘EpiScanpy’, following the basic steps described in the tutorial for scATAC-seq
461 ([https://github.com/colomemaria/epiScanpy/blob/master/docs/tutorials/Tutorial_Hackathon_B](https://github.com/colomemaria/epiScanpy/blob/master/docs/tutorials/Tutorial_Hackathon_Buenostro_2.html)
462 [uenostro_2.html](https://github.com/colomemaria/epiScanpy/blob/master/docs/tutorials/Tutorial_Hackathon_Buenostro_2.html)) with the same 50kbp matrices used for ChromSCape. We extracted from

463 each method the matrix of reduced feature space, and used hierarchical clustering with Pearson's
464 dissimilarity as the distance metric and Ward's method for linkage. The adjusted Rand's
465 index (ARI), a widely-used measure to quantify clustering accuracy, was calculated for each
466 method using R package 'mclust'²⁸, taking samples of origin as 'true' clusters.

467 In addition, two softwares with graphic interface, dedicated to the analysis of single-cell data,
468 'scOrange'¹³ and 'SCRAT'¹⁴, were also tested on the same set of cells both with 'default'
469 parameters and manually 'optimized' parameters. 'default' for scOrange corresponds to using
470 the template called 'Loading data from 10X protocols', a workflow meant for analyzing
471 scRNA-seq of bone marrow cells, replacing the input by our matrices of selected cells in 50kbp
472 bins. The 'optimized' workflow is available at
473 www.github.com/vallotlab/ChromSCape_benchmarking and can be opened with the
474 'scOrange' software. For 'SCRAT', we found that the 'optimized' counting method
475 corresponded to counting signal within peaks called on the 'pseudo-bulk' (see above).

476 In order to be able to compare the distinct methods, ChromSCape was first run on the raw count
477 matrices and a set of 1287 cells passing the quality control thresholds were selected to be used
478 as input for all methods. As the number of cells in each sample was unbalanced (e.g. the raw
479 MDA-MB-468 containing n=3,382 cells while others have a maximum of n=456 cells), 500
480 cells from MDA-MB-468 were randomly sub-sampled using ChromSCape 'Perform
481 Subsampling' option. We removed from the analysis the segments corresponding to known
482 amplifications and homozygous loss of DNA of the Triple Negative Breast Cancer cell line
483 MDA-MB-468, corresponding to a total of 77Mbp, previously found by analyzing the input of
484 bulk ChIP-seq of the same cells (see Supplementary Note 2).

485

486 **Code availability**

487 Source code, guidelines for installation and use of the application are provided at
488 <https://github.com/vallotlab/ChromSCape>. A docker container containing the application and
489 it's dependencies is available on DockerHub (pacomito/chromscape:v0.0.9001), instructions on
490 how to launch it are available on the github page. Codes for the benchmark of 'SnapATAC',
491 'CisTopic', 'Cusanovich2018' and 'EpiScanpy' are available at
492 https://github.com/vallotlab/ChromSCape_benchmarking.

493

494 **Data availability**

495 In this study, in addition of using publicly available single-cell datasets, we produced
496 H3K27me3 scChIP-seq data for MDA-MB-468 sample and a new replicate of untreated HBCx-
497 95. The sequencing data that support the findings of this study have been deposited in the
498 National Center for Biotechnology Information Gene Expression Omnibus (GEO) and are
499 accessible through the GEO Series accession number [GSE152502](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152502)
500 [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152502>]. All other relevant data
501 are available from the corresponding author on request.

502

503

504 **References**

- 505 1. Grosselin, K. *et al.* High-throughput single-cell ChIP-seq identifies heterogeneity of
506 chromatin states in breast cancer. *Nat. Genet.* 51, 1060–1066 (2019).
- 507 2. Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin
508 state. *Nature Biotechnology* 33, 1165–1172 (2015).
- 509 3. Harada, A. *et al.* A chromatin integration labelling method enables epigenomic profiling
510 with lower input. *Nat. Cell Biol.* 21, 287–296 (2019).
- 511 4. Ku, W. L. *et al.* Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile
512 histone modification. *Nat. Methods* 16, 323–325 (2019).
- 513 5. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and
514 single cells. *Nat Commun* 10, 1930 (2019).
- 515 6. Chen, X., Miragaia, R. J., Natarajan, K. N. & Teichmann, S. A. A rapid and robust method
516 for single cell chromatin accessibility profiling. *Nature Communications* 9, 5345 (2018).
- 517 7. Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by
518 combinatorial cellular indexing. *Science* 348, 910–914 (2015).
- 519 8. Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell
520 chromatin accessibility. *Nat. Biotechnol.* 37, 916–924 (2019).
- 521 9. Chen, H. *et al.* Assessment of computational methods for the analysis of single-cell ATAC-
522 seq data. *Genome Biol.* 20, 241 (2019).
- 523 10. Fang, R. *et al.* Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-
524 Regulatory Elements in Rare Cell Types. Preprint available at
525 <https://www.biorxiv.org/content/10.1101/615179v3> (2019).
- 526 11. Bravo González-Blas, C. *et al.* cisTopic: cis-regulatory topic modeling on single-cell
527 ATAC-seq data. *Nat. Methods* 16, 397–400 (2019). Preprint available at

- 528 12. Cusanovich, D. A. *et al.* The cis-regulatory dynamics of embryonic development at single-
529 cell resolution. *Nature* 555, 538–542 (2018).
- 530 13. Stražar, M. *et al.* scOrange-a tool for hands-on training of concepts from single-cell data
531 analytics. *Bioinformatics* 35, i4–i12 (2019).
- 532 14. Ji, Z., Zhou, W. & Ji, H. Single-cell regulome data analysis by SCRAT. *Bioinformatics* 33,
533 2930–2932 (2017).
- 534 15. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing,
535 quality control, normalization and visualization of single-cell RNA-seq data in R.
536 *Bioinformatics* 33, 1179–1186 (2017).
- 537 16. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level
538 analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* 5, 2122 (2016).
- 539 17. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput.*
540 *Biol.* 9, e1003118 (2013).
- 541 18. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell
542 RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat.*
543 *Biotechnol.* 36, 421–427 (2018).
- 544 19. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with
545 confidence assessments and item tracking. *Bioinformatics* 26, 1572–1573 (2010).
- 546 20. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine*
547 *Learning Research* 9, 2579–2605 (2008).
- 548 21. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and
549 Projection for Dimension Reduction. Preprint available at <https://arxiv.org/abs/1802.03426>
550 (2018).
- 551 22. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–
552 2079 (2009).

- 553 23. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
554 features. *Bioinformatics* 26, 841–842 (2010).
- 555 24. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137 (2008).
- 556 25. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for
557 differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140
558 (2010).
- 559 26. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and
560 Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B*
561 *(Methodological)* 57, 289–300 (1995).
- 562 27. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for
563 interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–
564 15550 (2005).
- 565 28. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: Clustering, Classification
566 and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal* 8, 289–317
567 (2016).
- 568 29. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human
569 hematopoiesis and leukemia evolution. *Nat. Genet.* 48, 1193–1203 (2016).
- 570 30. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory
571 variation. *Nature* 523, 486–490 (2015).
- 572 31. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring
573 transcription-factor-associated accessibility from single-cell epigenomic data. *Nat.*
574 *Methods* 14, 975–978 (2017).
- 575 32. Danese, A., Richter, M. L., Fischer, D. S., Theis, F. J. & Colomé-Tatché, M. EpiScanpy:
576 integrated single-cell epigenomic analysis. Preprint available at
577 <https://www.biorxiv.org/content/10.1101/648097v1> (2019).

579 **End Notes**

580 **Funding**

581 This research project was supported by the ATIP Avenir program, by Plan Cancer and by the
582 SiRIC-Curie program SiRIC Grants #INCa-DGOS- 4654 and #INCa-DGOS-Inserm_12554.

583

584 **Authors contribution**

585 PP, PK and CV wrote code for ChromSCape and performed data analysis. JM performed
586 scChIP-seq experiments on PDX and breast cancer cell line. PP, MD and NS processed raw
587 scChIP-seq datasets into count matrices. PP and CV wrote the manuscript. CV conceptualized
588 and supervised this work.

589

590 **Competing interests**

591 Licenses have been filed on some aspects of this work by Institut Curie and CNRS; contributors
592 may receive payments related to exploitation of these licenses under their employer's rewards
593 to inventor scheme.

594

595

596 **Legends**

597 **Figure 1. Representation of ChromSCape workflow**

598 Users upload single-cell epigenomic data formatted as count matrices, single-cell BAM or
599 single-cell BED files to start the analysis. The application includes Quality Control (QC),
600 Classification and Interpretation tools. The user can save plots and tables in png, pdf or csv
601 formats, and R analysis objects in RData format.

602

603 **Figure 2. Benchmarking single-cell epigenomic tools with an *in-silico* mix of H3K27me3**

604 **scChIP-seq.** The mix is composed of human cells from an untreated PDX (HBCx-22), human
605 T cells (Jurkat) and B cells (Ramos) taken from ¹ and from a TNBC cell line (MDA-MB-468).

606 (a) UMAP plots obtained with ChromSCape colored according to cluster and sample of origin.

607 Adjusted Random Indexes (ARI) is indicated above the plot. (b) UMAP plots colored according
608 to cluster and sample of origin with other single-cell epigenomic analysis methods:

609 *Cusanovich2018*, SnapATAC, CisTopic and EpiScanpy. Adjusted Random Indexes (ARI) are

610 indicated above the plots. (c) Snapshots from scOrange and SCRAT applications. PCA and t-

611 SNE representations from scOrange and SCRAT respectively, using default parameters or after

612 manually optimizing parameters.

613

614 **Figure 3. ChromSCape identifies immune cell populations from scATAC-seq datasets.** (a)

615 t-SNE representations after correlation filtering (n=1309 cells), points are colored according to

616 sample of origin (left) or ChromSCape-determined cluster ($k=5$) (right). The GM12878 and

617 K562 samples contained respectively 4 and 6 replicates. (b) Assignment scores for each

618 sample/cluster pair for the analysis with all samples. (c-d) As in (a) and (b) for the analysis with

619 only AML, LSC, monocyte, LMPP & HL60 cells (n=347 cells).

620

621

622 **Figure 4. ChromSCape deconvolves epigenomic landscapes within the tumor micro-**
623 **environment.** Cells belong to samples HBCx-22, HBCx-22-TamR, HBCx-95, HBCx-95-
624 CapaR PDX¹. PCA and t-SNE plots are colored according to the amount of uniquely mapped
625 reads per cell (a) and to sample of origin (b). (c) t-SNE representations after correlation filtering
626 ($n = 903$ cells), colored by cluster or sample of origin. (d) Hierarchical clustering and
627 corresponding heatmap of cell-to-cell consensus clustering scores cells portioning the dataset
628 into $k = 4$ clusters. Consensus score ranges from 0 (white: never clustered together) to 1 (dark
629 blue: always clustered together). Cluster membership is color-coded above the heatmap. (e)
630 Table of cluster memberships. P-value column results from Pearson's Chi-squared goodness of
631 fit test without correction, checking if the observed distribution of samples in each cluster
632 differs from random distribution. Source data are provided as a Source Data file. (f) t-SNE
633 representation of scChIP-seq datasets, points are colored according to H3K27me3 enrichment
634 signals in each cell for genes located within depleted regions in C1 to C4, respectively *Eln*,
635 *Bcar1*, *Nrros* and *Rap1gap2*. The adjusted p-values and log₂FC of the associated regions are
636 indicated above each plot. (g) Barplot of differentially enriched regions identified by Wilcoxon
637 signed-rank test. Genomic regions were considered enriched (red) or depleted (green) in
638 H3K27me3 if the adjusted p-values were lower than 0.01 and the absolute fold change greater
639 than 1. (h) Barplot displaying the $-\log_{10}$ of adjusted p-values from pathway analysis for cells
640 of cluster C2 compared to all other cells in depleted loci. Only the top 15 significant gene sets,
641 ranked by adjusted p-values, are indicated.

642

Fig. 1

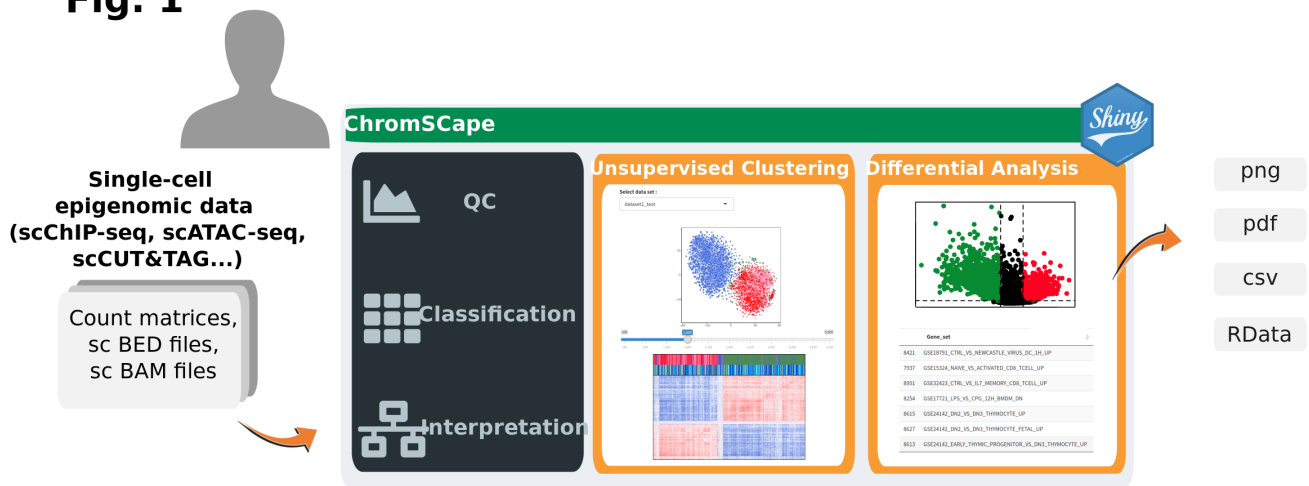


Fig. 2

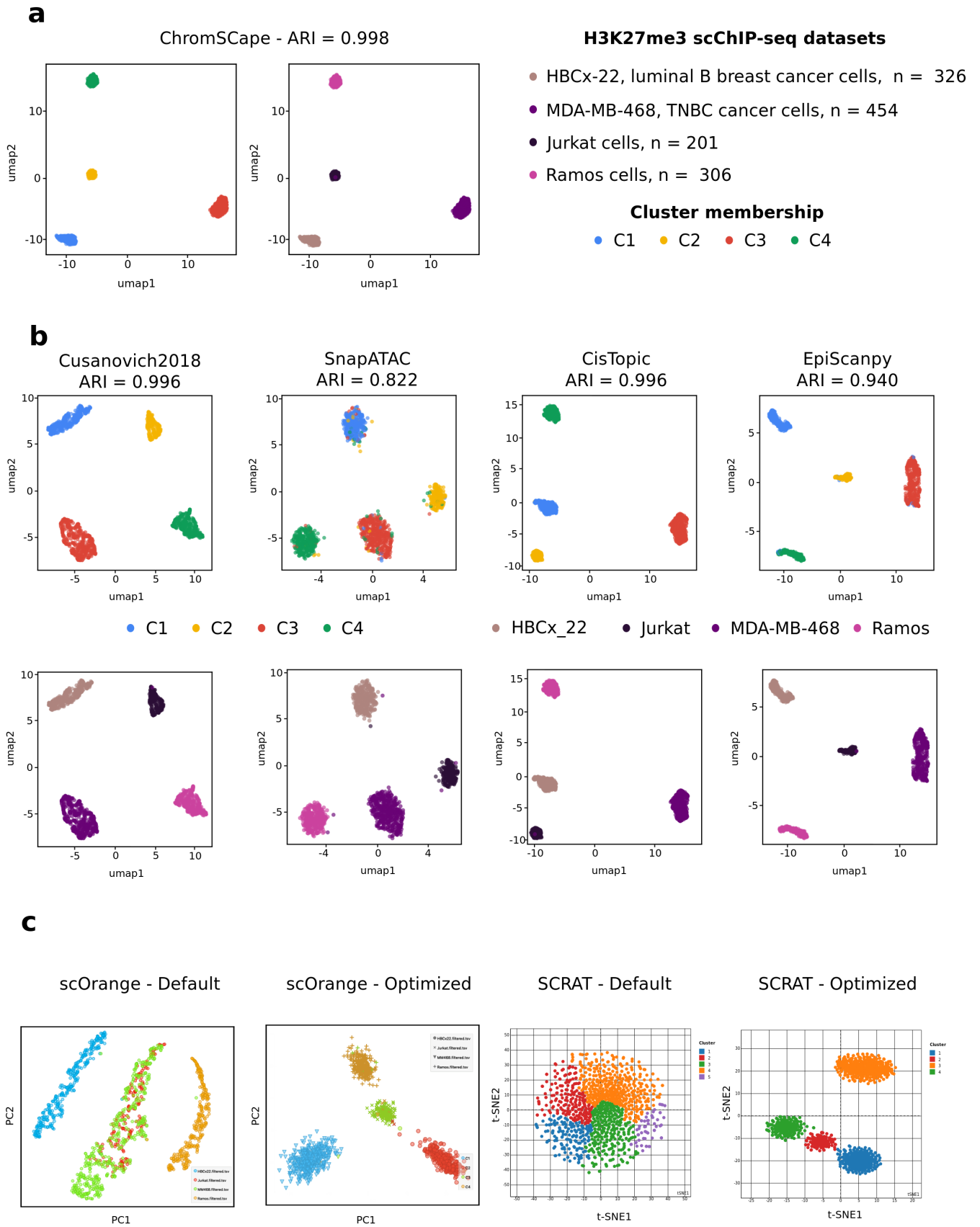


Fig. 3

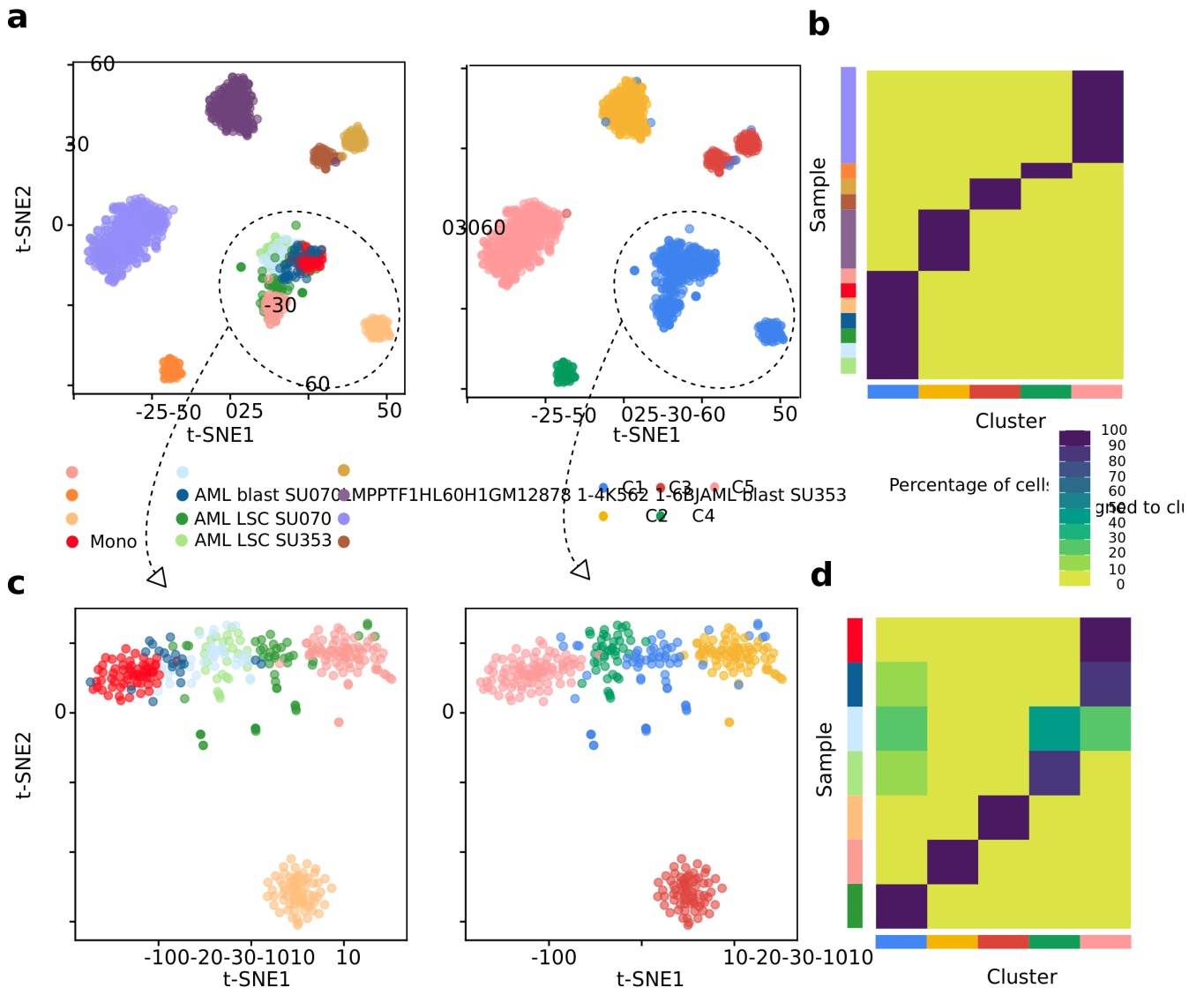


Fig. 4

