Characterizing epigenetic heterogeneity at the cellular level is a critical problem in the modern genomics era. Assays such as single cell ATAC-seq (scATAC-seq) offer an opportunity to interrogate cellular level epigenetic heterogeneity through patterns of variability in open chromatin. However, these assays exhibit technical variability that complicates clear classification and cell type identification in heterogeneous populations. We present scABC, an R package for the unsupervised clustering of single-cell epigenetic data, to classify scATAC-seq data and discover regions of open chromatin specific to cell identity.
Recent advances in single cell technologies such as scATAC-seq\textsuperscript{1,2} and scChIP-seq\textsuperscript{3} have expanded our understanding of epigenetic heterogeneity at the single cell level. However, datasets arising from such technologies are difficult to analyze due to the inherent sparsity. In particular, consider scATAC-seq, designed to interrogate open chromatin in single cells. Open sites in a diploid genome have at most 2 chances to be captured through the assay and only a few thousand distinct reads are generated per cell, resulting in a very low chance that a particular site is captured by the assay. Consequently, it is difficult to determine whether a region is absent in an individual cell due to the lack of openness or due to the sparse nature of data. This creates a challenging task in delineating distinct sub-populations, as only a few genomic regions will have overlapping reads in a large number of cells. To avoid this issue, many studies perform FACS sorting to identify subpopulations, followed by bulk sequencing to determine genomic regions of interest and guide the single-cell analysis. If the population is unknown or marker genes are unavailable, then sub-population specific analysis becomes impractical with these techniques.

To combat these challenges and allow for the de novo classification of individual cells by their epigenetic signatures, we present a statistical method for the unsupervised clustering of scATAC-seq data, named single cell Accessibility Based Clustering (scABC). In contrast to previous works\textsuperscript{2,4} that demand predefined accessible chromatin sites, our procedure relies solely on the patterns of read counts within genomic regions to cluster cells. It requires two inputs: the individual single cell mapped read files and the full set of called peaks (which can be obtained from the union of all of the individual cells without the need for additional experiments). We apply our method to publicly available scATAC-seq data\textsuperscript{1,2,4}, as well as a true biological mixture to show that our approach can cluster cells with similar epigenetic patterns and identify accessible regions specific to each cluster. We further demonstrate that the cluster specific accessible regions determined by scABC have functional meaning and are capable of determining cellular identity. In particular, we show that these cluster specific accessible regions are enriched for transcription factor motifs known to be specific to each subpopulation and that, through association with scRNA-seq data, they can lead to the identification of subpopulation specific gene expression.

**Results**

**The scABC algorithm.** First, we briefly describe our algorithm and the intuition behind it (Fig. 1a). To tackle the problem of sparsity, we noted that cells with higher sequencing coverage should be more reliable since important open regions are less likely to be missed by random chance. Therefore, scABC first weighs cells by (a nonlinear transformation of) the number of distinct reads within peak backgrounds and then applies a weighted K-medoids clustering\textsuperscript{5} to partition the cells into distinct groups (see Methods for details). scABC uses the ranked peaks in each cell to perform the clustering rather than the raw counts to prevent bias from highly over-represented regions. We found that this usually sufficient to cluster most cells, but a few problematic cells seem to be misclassified. To improve the classification, we calculate landmarks for each cluster. These landmarks depict prototypical cells from each cluster and are characterized by the highest represented peaks in each cluster, which we should trust more than the noisy low-represented peaks. scABC finally clusters the cells by assignment to the closest landmark based on the Spearman correlation (Fig. 1b). With the cluster assignments we can then test whether each accessible region is specific to a particular cluster, using an empirical Bayes regression based hypothesis testing procedure to obtain peaks specific to each cluster (Fig. 1c, Methods).

**Performance evaluation using in silico mixture of cells.** To test our method, we constructed an in silico mixture of 966 cells from 6 established cell lines, previously presented in Buenrostro et al.\textsuperscript{1}

**Fig. 1** The scABC framework for unsupervised clustering of scATAC-seq data. a Overview of scABC pipeline. scABC constructs a matrix of read counts over peaks, then weights cells by sample depth and applies a weighted K-medoids clustering. The clustering defines a set of K landmarks, which are then used to reassign cells to clusters. b Assignment of cells to landmarks by Spearman correlation, where each cell is highly correlated with just one landmark. The similarity measure used above is defined as the Spearman correlation of cells to landmarks, normalized by the mean of the absolute values across all landmarks for every cell. This allows us to better visualize the relative correlation across all cells. c Accessibility of peaks across all cells. The vast majority of peaks tend to be either common or cluster specific, allowing us to define cluster specific peaks.
We designed a test to systematically assess scABC sensitivity. For each cell line, we equally divided its cells into two groups and replaced a fraction of peaks in one group using another cell line. Applying scABC to these two groups, we achieve successful classifications when at least 50–70% of peaks are identical between the groups (Supplementary Fig. 6). In later sections, we will evaluate the sensitivity of scABC on real mixtures that have similar sub-populations.

We next investigated whether the cluster specific peaks obtained by scABC are able to define cell identity (Supplementary Fig. 7). These peaks contain both narrow and broad regions, as defined by MACS2. In principle, narrow peaks better capture TF binding sites. To measure the enrichment of TF motifs in individual cells, we applied chromVAR to narrow peaks with scABC defined p-value < 10⁻⁶, named cluster specific narrow peaks. This cutoff was chosen because it approximately equals the Bonferroni corrected cutoff for a family wise error rate of 0.05. The full details to reproduce the chromVAR results are outlined in scABC vignettes, available online with the software package (see Code availability for details). chromVAR calculates deviations, essentially z-scores for TF motif enrichment that are normalized for background accessibility and other biases such as GC content. We found that the most active TFs are typically specific to one or two clusters, identifying active TFs in every cell type (Fig. 2a). Some of these TFs were previously shown to be context-specific, for instance, NFKB2 in GM12878 cells, SPI1 in HL-60 cells, GATA1::TAL1 in K562 cells, and FOSS in BJ cells. It is important to note that TFs with similar DNA-binding motifs show similar motif enrichments. Therefore, POU motifs that are enriched in H1 can demonstrate the activity of POU5F1, the core regulator of human embryonic stem cell self-renewal. We observe that BJ specific TFs seem to be better distinguished than other TFs. Because BJ cells are dissimilar to any of the other cell lines (Fig. 1b) and have by far the highest number of cluster

![Fig. 2 Cluster specific peaks determined by scABC shed light on cell identity.](https://example.com/fig2)

a. Application of chromVAR to the cluster specific narrow peaks allows for the identification of cluster specific transcription factor binding motifs. chromVAR calculated deviations are shown for the top twenty most variable transcription factor binding motifs. b. Cluster-specific open promoters distinguish expression. Shown are the densities of the average log gene expression values in genes with either a K562-specific open promoter, HL60-specific open promoter, or non-specific promoter (neither) in K562 cells (left) or HL60 cells (right), with each plot normalized to have total area equal to one. c. Integration of scATAC-seq and scRNA-seq enables clear delineation of cell identity. scABC applied to scATAC-seq identified genes with cluster specific open promoters for K562 and HL60 cells. These genes were then used for Principal Component Analysis (PCA) of 42 K562 and 54 HL-60 cells (right) and compared to PCA of all genes (left).
specific peaks (Fig. 1c), this is not unexpected. We also applied chromVAR to the full set of narrow peaks and found comparable results (Supplementary Fig. 8), indicating that the cluster specific peaks are responsible for the majority of the variation while comprising <15% of all narrow peaks.

In contrast to narrow peaks, broad regions are more suited to demonstrate functional DNA elements such as promoters and enhancers. We hypothesized that cluster specific broad peaks overlapping gene promoters have functional significance and can help distinguish genes specific to a particular cell type. Specifically, we expect that genes with cell type specific open promoters will have, on average, higher expression in that cell type versus the other cell types in the population. To evaluate this hypothesis, we took 42 K562 and 54 HL-60 deeply sequenced scABC experiments (Supplementary Note). We defined a gene to have a cell type specific open promoter if any open peak with an scABC p-value of <10^{-6} overlapped more than 400 base pairs in the gene promoter, defined as the region 5 kb upstream of the primary FANTOM5 TSS.

We first confirmed our hypothesis that genes with cell type specific open promoters tend to be higher in that cell type, compared to other genes (Fig. 2b). We next clustered the corresponding gene expression data (in transcripts per million, named TPM) using both all genes and only those genes with cell type specific open promoters in K562 and HL-60 cells (as shown in Fig. 2c). After normalization for batch effects and clustering based on all genes did not clearly separate the two cell types in the first two principal components. When genes associated with cell-type specific open promoters were employed, the separation became extremely obvious. Similar patterns were observed when using t-SNE plots (Supplementary Fig. 9). This verifies our hypothesis that cluster specific broad peaks shed light on functional significance outside of motif enrichment.

Performance evaluation on experimental mixtures. In addition to the in silico cell line mixture, we examined the capability of scABC in classifying three heterogeneous populations. We first applied scABC to experimental mixtures of GM12878 and HEK293T cells as well as GM12878 and HL-60 cells. These experiments, cells were processed in a single batch for each mixture. In both cases, clear separation between the two cell lines were achieved (Supplementary Figs. 10 and 11) that, due to the experimental design, cannot be explained by batch effects. Although we correctly classified these cell lines, they are from fairly distinct origins and easy to separate.

To tackle a more difficult problem, we return to the analysis of the GM12878 cell line. Recall that when we intentionally set the number of clusters equal to 4 we found 2 slightly similar clusters. These results were consistent when we set K = 2 (Supplementary Fig. 12). We hypothesized that these small variations may suggest heterogeneity in the GM12878 cell line. We observed that one cluster is enriched for NF-kB motifs, such as NFKB2, REL, and RELA, and this may be an indication of transcription factor heterogeneity. The nuclear localization of NF-kB was previously shown to dynamically change and cause temporal variations in transcription factor expression, which may explain this heterogeneity. Previous studies have also suggested that cellular variability in GM12878 may be driven by NF-kB heterogeneity. These findings are consistent with our clustering results, but, we cannot further confirm them due to incomplete biological knowledge of GM12878 cell heterogeneity.

Application to a heterogeneous biological population. For a reliable assessment of our method, we generated a heterogeneous biological population of cells that arise from the same origin. Specifically, we used the hanging drop technique to form embryoid bodies (EBs) from mouse embryonic stem cells (mESCs). We next differentiated EBs using retinoic acid (RA) treatment and performed scATAC-seq on day 4 of the development (Methods). We generated a single 96-well plate and obtained 95 cells that pass quality control (Supplementary Note).

It is well known that RA-treated mESCs are induced to differentiate into neuronal cell types. However, when three-dimensional EBs are treated with RA, previous studies have suggested that the outer layer of EBs expresses marker genes that are characteristic of visceral endoderm cells during early mouse development. We therefore hypothesized that the RA-treated EBs are a heterogeneous mixture, consisting of both visceral endoderm and neural ectoderm cells. To confirm this heterogeneity in terms of chromatin accessibility, we applied scABC to the 95 cells and obtained K = 2 clusters (Supplementary Fig. 3) with well separated landmarks (Fig. 3a) and cluster specific peaks (Fig. 3b). We next ran chromVAR on the cluster specific narrow peaks and found that almost all TFs are specific to one cluster (Fig. 3c). The majority of TFs active in cluster 1 play key roles in neural development, including GLI1/2, LBX1, LMX1A, MNX1, NEUROG2, NKX6-1/2, UNCX1, VAX1, and Pou factors. The high activity of LHX3/9 in cluster 1 may be related to LHX3/4 (because of their similar motifs), which have been shown to function in the development of mouse motor neuron. In contrast to cluster 1, TFs specific to cluster 2 are essential for visceral endoderm differentiation, such as GATA factors, HNF1A/B, and the AP-1 family (i.e., JUN and FOS motifs). The TF enrichment analysis suggests that scABC clearly distinguishes neuroectoderm (67 cells) from visceral endoderm (28 cells), two sub-populations with the same origin (mEB) in early embryonic development.

Since chromVAR only reflects motif enrichment and cannot distinguish TFs with similar DNA-binding motifs, we next sought to narrow down the list of TFs using bulk RNA-seq data at day 6 of the development (the closest publicly available RNA-seq to day 4). We first used TOMTOM too to identify TFs with similar motifs to those enriched in cluster 1 and 2 (q-value ≤ 10^{-4}) and then selected a subset of them that are highly expressed in bulk RNA-seq data (FPKM ≥ 10). Interestingly, the majority of motifs that were found enriched in cluster 1 and 2 are associated with expressed TFs. Specifically, the expressed genes BHLHE22, SHOX2, GBX2, HOXB2/3/5, LHX4, MNX1, NEUROG2, NKX6-1/2, OLG1/2, and POUSF1 to POUSF4 are active (in terms of motif enrichment) in cluster 1 while FOS, FOXB1, GATA4, JDP2, JUN, and JUND are active in cluster 2. These findings indicate that both neuroectoderm and visceral endoderm sub-populations are active at RNA level.

scABC characterizes the leukemic evolution. To further extend the evaluation of scABC performance, we tested its ability for detecting developmental stages of cancer evolution. Corces et al. sequenced individual monocytes and lymphoid-primed multi-potent progenitors (LMPP) from healthy donors and leukemia stem cells (LSC) and leukemic blast cells (blast) from donors with acute myeloid leukemia. Notably, this dataset is extremely sparse compared to the in silico mixture of 6 cell lines (Supplementary Fig. 1). We applied scABC followed by chromVAR to the combined mixture of the 390 cells that passed quality control (Supplementary Note). Our method detected K = 2 clusters, which resulted in a clear separation of the cells into a monocyte dominated cluster and a LMPP dominated cluster with blasts predominantly clustered with monocytes and LSCs mainly clustered with LMPPs (Supplementary Figs. 3 and 13, and Supplementary Table 5, and Supplementary Note). When using more
Intervals (100 kb, binned using the software csaw), as well as marks) using read counts in peaks and binned counts over long intervals notably increased the number of misclassifications for both methods (Supplementary Tables 9 and 10), suggesting that peaks better reflect chromatin accessibility. Hence, we used peaks for the remaining method comparisons.

We next compared scABC to SC3 (a clustering method designed specifically for scRNA-seq) and a community structure clustering method based on the infomap algorithm46,47. Applying these methods to the in silico mixture, SC3 did not distinguish the BJ cells from HL60 cells, despite our results indicating that BJ cells are well separated from all other cell types (Supplementary Table 11). On the other hand, the community structure clustering method seemed little better than random assignment (Supplementary Table 12). These results indicate to us that scRNA-seq clustering methods are unlikely to easily generalize to scATAC-seq, which we believe is due to the extreme sparsity of scATAC-seq data.

To clarify the differences between simple K-medioids (with the best performance among the alternative methods) and scABC, we downsampled each cell line and found that scABC is able to...
identify small subpopulations (Supplementary Fig. 15). We next applied K-medoids to the RA-treated EB cells and found that the sub-populations identified were not biologically meaningful when we examined TF enrichment (Supplementary Fig. 16 and Table 13, see the previous section for scABC clustering results).

To evaluate the performance of scABC’s method of determining cluster specific peaks, we used peaks differentially open in the respective bulk data as a gold standard and compared scABC to an existing method for identifying differentially expressed genes in single cell RNA-seq, SCDE53,54 (Supplementary Note). We found that the majority of cluster specific peaks identified by scABC are differentially open in the respective bulk data and the overlap was much larger than the differentially expressed peaks of SCDE (Supplementary Figs. 17 and 18). We also observed that SCDE calculated cluster specific peaks are not well separated (Supplementary Fig. 19), compared to scABC (Supplementary Figs. 10 and 11). We note that since scATAC-seq data tends to be sparser and have lower read counts than scRNA-seq data, it is not surprising that methods developed for scRNA-seq data, such as SCDE, may not easily generalize to scATAC-seq data.

Discussion

In summary, we developed scABC for the unsupervised clustering and identification of cluster specific peaks for single cell epigenetic data. We showed that scABC can be applied to scATAC-seq data of complex mixtures to deconvolve the underlying population structure. We should note that in cases where the population lies in a continuum, scABC will not be able to separate the population. In our experience, this is usually indicated by a continuously increasing gap statistic. In such cases other tools such as graph embedding49 or k-mer analysis8,50 may be more appropriate.

We showed that the scABC identifies informative peaks for downstream analysis. Since scABC only uses the read counts within peaks to identify informative peaks, further analysis on the content of the peaks can be done in an unbiased manner while increasing the signal to noise ratio. For example, we showed that scABC in conjunction with chromVAR identifies the drivers of cellular heterogeneity in developmental dynamics in the context of retinoic acid induction. In another example, we showed that cell type specific open promoters can better identify cell type specific expression.

Methods

Unsupervised clustering of scATAC-seq data. The clustering algorithm of scABC can be broken down into three steps.

Weighted K-medoids clustering: Cells with low sequencing depth are noisy and can negatively impact the clustering result. We implement a weighted version of the K-medoids clustering algorithm, where cells with lower sequencing depth are given smaller weight. Let \( h_i \) denote a measure of relative sequencing depth for cell \( i \). For each clustering step, \( h_i \) is multiplied by \( 1 - h_i \) to give a measure of cell quality.

\[
\begin{align*}
\text{w}_i &= 1 + \exp\left(-\frac{1}{(h_i - c)/(\lambda)}\right),
\end{align*}
\]

where \( c \) and \( \lambda \) are tuning parameters. As defaults, we use the median of the background and 0.1, respectively. We found that the performance of the clustering is robust to a wide range of \( [c, \lambda] \) (Supplementary Table 14).

Let \( Y_i \) denote the read counts within peaks for cell \( i \) (dimension of \( Y_i \) is equal to the number of input peaks), \( K \) the number of clusters, \( C \) the cluster assignment, and \( i_t \) the medoid for cluster \( k_t \), i.e. a cell used as the cluster center. The clustering assignment is given by the solution to

\[
\min_{c, \lambda, Y_{ik_t}} \sum_{k=1}^{K} \sum_{i=1}^{N} w_i d(Y_i, Y_{ik_t}),
\]

where \( d(\cdot) \) is in general the dissimilarity between a pair of samples. We use 1–Spearman’s rank correlation as the dissimilarity measure, and refer to the Spearman rank correlation as the similarity measure. The problem above is solved by the Partitioning Around Medoids (PAM) algorithm51 as implemented in the R package WeightedCluster2.

Landmarks: We sum the reads across the cells within a cluster and select the \( P \) peaks with the highest read counts to obtain the landmark for each cluster identified in the previous step. As a default we set \( P = 2000 \).

Re-clustering using landmarks: To refine the clustering results, we re-cluster the cells by assigning each cell to the landmark with the highest Spearman’s rank correlation using the union of all landmark peaks.

The weighted K-medoids algorithm requires the number of clusters \( K \) in advance. We determine \( K \) through the gap statistic52 with a few modifications to better capture the data structure of single cell experiments, particularly sparsity and cell heterogeneity (Supplementary Fig. 5, Supplementary Note).

Identification of cluster specific peaks. To find peaks that tend to be more open in one cluster than all others, we formulate the problem in a hypothesis testing framework. We perform the hypothesis testing on all peaks but the procedure is applicable to any subset of peaks, such as narrow or broad peaks. We first introduce our statistical models and then focus on the strategy.

Model assumption: Let \( K \) denote the number of clusters, \( R \) the total number of peaks, \( y_{ri} \) the read counts for peak \( r \) in cell \( i \), and \( x_{rk} \) the cluster membership for cell \( i \) with \( x_{rk} = 1 \) if cell \( i \) belongs to cluster \( k \) and \( x_{rk} = 0 \) otherwise. We assume that \( y_{ri} \) follows a Poisson distribution with mean \( \mu_{ri} \)

\[
\mu_{ri} = h_i q_{ri},
\]

\[
\log q_{ri} = \beta_0 + \sum_{k=1}^{K} x_{rk} \beta_k,
\]

where \( \beta_0 \) is the intercept and the coefficients \( \beta_k \) exhibit the effect of the cluster membership on peak \( r \). We assume normal priors on the cluster membership effects.

Empirical prior estimate: The normal prior enables empirical Bayes shrinkage on \( \beta_0 \), and stabilizes the noisy estimate when the read counts are low53. To obtain a robust empirical prior estimate \( \hat{\beta}_k \), we adopt the quantile matching method proposed in DESeq253. In particular, we first fit a model without the intercept \( \beta_0 \) and without the normal prior to obtain the maximum likelihood estimate (MLE) \( \beta_mle \). Let \( \theta = \sum_{k=1}^{K} \beta_mle \) and let \( \hat{\beta}_k \) denote the vector \( \{\beta_0, \beta_1, \ldots, \beta_k\} \). We define the empirical cdf of \( |\beta_mle - \beta_k| \), with \( \Phi(\cdot; |\gamma|) \) the cdf for the standard normal quantile. The empirical prior estimate for the standard deviation is calculated as

\[
\sigma_k = \frac{\Phi(\cdot; \beta_mle - \beta_k)}{\sigma_{\text{prior}}}, \quad \forall k = 1, \ldots, K.
\]

We set \( \sigma = 0.05 \) in practice. Details for computing \( \sigma_k \) are described in the Supplementary Note.

Hypothesis testing: Suppose \( \Gamma_k = \{1, \ldots, K\} \) represent the set \( \{1, \ldots, K\} \) for the \( k \)-th component. To test whether peak \( r \) is specific to cluster \( k \), we consider

The null hypothesis \( H_0_k : \beta_{rk} \leq \beta_{rk}' \), for some \( k' \in \Gamma_k \)

The alternative hypothesis \( H_1_k : \beta_{rk} > \beta_{rk}' \), for all \( k' \in \Gamma_k \).

Following the intersection-union test54, the null hypothesis can be broken into \( K - 1 \) simpler null hypotheses \( H_{0k} : \beta_{rk} \leq \beta_{rk}' \), \( k \in \Gamma_k \). For each null hypothesis, the Wald test statistics is \( \left(\beta_{rk} - \beta_{rk}'\right)/SE(\beta_{rk} - \beta_{rk}') \), where \( \beta \) is the maximum a posteriori (MAP) estimate for \( \beta \) and \( SE(\cdot) \) the MAP estimated standard error, which depends on both the observed data and prior estimates. The rejection region for \( H_{0k} \), with size \( \alpha \)

\[
\frac{\beta_{rk} - \beta_{rk}'}{SE(\beta_{rk} - \beta_{rk}')} > z_{\alpha},
\]

and the rejection region for \( H_k \) with level \( \alpha \)

\[
\frac{\beta_{rk} - \beta_{rk}'}{SE(\beta_{rk} - \beta_{rk}')} > z_{\alpha}.
\]

Details for computing \( \beta \) and the standard errors are illustrated in the Supplementary Note. We finally compute the \( p \)-value for \( H_{0k} \) as \( \max\{p_{rk}, k' \in \Gamma_k \} \), with \( p_{rk} \) indicating the \( p \)-value for \( H_{0k} \).

Experimental design of RA-treated mESC differentiation. Cell culture: Mouse ES cell lines R1 were obtained from ATCC. The mESCs were first expanded on an MEF feeder layer previously irradiated. Then, subculturing was carried out on 0.1% bovine gelatin-coated tissue culture plates. Cells were propagated in mESC medium.
consisting of Knockout DMEM supplemented with 15% Knockout Serum Replacement, 100 μM nonessential amino acids, 0.5 mM beta-mercaptoethanol, 2 mM GlutaMax, and 100 μM Penicillin-Streptomycin with the addition of 1,000 μM of LIF (ESGRO, Millipore).

Cell differentiation: mESCs were differentiated using the hanging drop method. Trypsinized cells were suspended in differentiation medium (mESC containing 0.5 M cell strainer (pluriSelect) to remove undissociated EBs. Before loading, the cells were washed three times in C1 DNA Seq Cell Wash Buffer (Fluidigm). In total 9 μL at a concentration of 400 cells/μL were combined with C1 Cell Suspension Reagent at a ratio of 3:2 and 10 μL of this cell mix was loaded on to a Leica CTR 6000 microscope for imaging, followed by Tn5 transposition and primary 8 cycles of PCR using the "ATASeq: Sample Prep (1861x/1862x/1863x)" scripts. The entire volume (3.5–5 μL) of the amplified transposed DNA was transferred to a 96-well plate containing 10 μL of C1 DNA Dilution Reagent. In the 96-well plate, harvested libraries were further amplified in 50 μL PCR (1.25 μL custom Nextera dual-index PCR primers in 1x NEBNext High-Fidelity PCR Master Mix) using the following PCR conditions: 72 °C for 5 min; 98 °C for 30 s; and total 14 cycles of: 98 °C for 10 s, 72 °C for 30 s, and 72 °C for 1 min. The PCR products were pooled together (4.8 mL) and the pooled library was purified on a single MiniElute PCR purification column (Qiagen) and eluted in 20 μL of Elution Buffer. Libraries were quantified using qPCR prior to sequencing using Illumina NextSeq 500 (paired-end 75 bp).

Code availability. The scATAC-seq data generated from RA-treated mESCs have been deposited in the Gene Expression Omnibus (GEO) under the accession number GSE107651. Other datasets used in this work are cited in the paper, with the accession codes provided in Supplementary Note.

Received: 4 December 2017 Accepted: 7 May 2018
Published online: 20 June 2018

References
1. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486–490 (2015).
2. Cusanovich, D. A. et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science 348, 910–914 (2015).
3. Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat. Biotechnol. 33, 1165–1172 (2015).
4. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat. Genet. 48, 1193–1203 (2016).
5. Studer, M. WeightedCluster library manual: a practical guide to creating weights for gene expression data. Nat. Methods 7, 1257–1274 (2010).
6. Li, H. et al. Integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat. Biotechnol. 26, 1293–1300 (2008).
7. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell genomic data. Nat. Methods 14, 575–583 (2017).
8. Smith, L. T., Hohaus, S., Gonzalez, D., Dziennis, S. & Tenen, D. PU.1 (Spi-1) and C/EBP alpha regulate the granulocyte colony-stimulating factor receptor promoter in myeloid cells. Blood 88, 1234–1247 (1996).
9. Gerstein, M. B. et al. Architecture of the human regulatory network derived from ENCODE data. Nature 489, 91–102 (2012).
10. Szyprytz, G., Feng, I., Zlinskiwi, T. V & Vileppontea, B. Similar gene expression pattern in senescent and hypertrophic-treated fibroblasts. J. Gerontol. A Biol. Sci. Med. Sci. 53, B438–B442 (1998).

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-04629-3 | www.nature.com/naturecommunications

ARTICLE

NATURE COMMUNICATIONS | 2018:9 2410

DOI: 10.1038/s41467-018-04629-3 | www.nature.com/naturecommunications

7
41. Wilkinson, D. G., Bhatt, S., Ryseck, R.-P. & Bravo, R. Tissue-specific expression of c-Jun and JunB during organogenesis in the mouse. *Development* **106**, 463–471 (1989).
42. Bailey, T. L. et al. Meme suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
43. Steidl, U. et al. Essential role of jun family transcription factors in PU.1 knockdown-induced leukemic stem cells. *Nat. Genet.* **38**, 1269–1277 (2006).
44. Lun, A. T. & Smyth, G. K. From reads to regions: a Bioconductor workflow to detect differential binding in ChIP-seq data. *F1000Res.* **4**, 1080 (2015).
45. Kiselev, V. Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483 (2017).
46. Rosvall, M. & Bergstrom, C. T. Mapping change in large networks. *PLoS ONE* **5**, e6894 (2010).
47. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955 (2017).
48. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
49. Pliner, H. et al. Chromatin accessibility dynamics of myogenesis at single cell resolution. Preprint at *Biorxiv* https://www.biorxiv.org/content/early/2017/06/26/155473 (2017).
50. de Boer, C. & Regev, A. Deciphering variance in epigenomic regulators by k-mer factorization. Preprint at *Biorxiv* https://www.biorxiv.org/content/early/2017/04/21/129247 (2017).
51. Rousseeuw, P. J. & Kaufman, L. *Finding Groups in Data.* (Wiley Online Library, Hoboken, NJ, 1990).
52. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc.: Ser. B* **63**, 411–423 (2001).
53. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
54. Casella, G. & Berger, R. L. *Statistical Inference.* (Duxbury, Pacific Grove, CA, 2002).
55. Wang, X. & Yang, P. In vitro differentiation of mouse embryonic stem (mES) cells using the hanging drop method. *J. Vis. Exp.* **17**, e825–e825 (2008).

**Acknowledgements**

This work was supported by grants R01HG007834, P50HG007735, and R01GM109836 from the National Institutes of Health (NIH). We thank Dr. Michael Snyder and members of Wong, Greenleaf, and Chang Labs for helpful suggestions.

**Author contributions**

M.Z., Z.L., T.D., and W.H.W. conceived the project. M.Z., Z.L., T.D., Z.D.; A.N.S. performed the analysis with inputs from W.J.G. and W.H.W.; X.C. generated the RA induction data; and M.Z., Z.L., and T.D. prepared scABC R package. All authors wrote the manuscript.

**Additional information**

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-04629-3.

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

**Reprints and permission** information is available online at http://npg.nature.com/reprintsandpermissions/

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

© The Author(s) 2018