nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-04629-3 OPEN

Unsupervised clustering and epigenetic classification of single cells

Mahdi Zamanighomi[1], Zhixiang Lin[1], Timothy Daley[1,2], Xi Chen[1,3], Zhana Duren[1], Alicia Schep[4,5], William J. Greenleaf[4,5,6] & Wing Hung Wong[1,3]

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.

[1] Department of Statistics, Stanford University, Stanford, CA 94305, USA. [2] Department of Bioengineering, Stanford University, Stanford, CA 94305, USA. [3] Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA. [4] Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA. [5] Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA 94305, USA. [6] Department of Applied Physics, Stanford University, Stanford, CA 94305, USA. These authors contributed equally: Mahdi Zamanighomi, Zhixiang Lin, Timothy Daley. Correspondence and requests for materials should be addressed to W.H.W. (email: whwong@stanford.edu)

Recent advances in single cell technologies such as scATAC-seq[1,2] and scChIP-seq[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 cells, 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[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[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 weights cells by (a nonlinear transformation of) the number of distinct reads within peak backgrounds and then applies a weighted *K*-medoids clustering[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.[1]

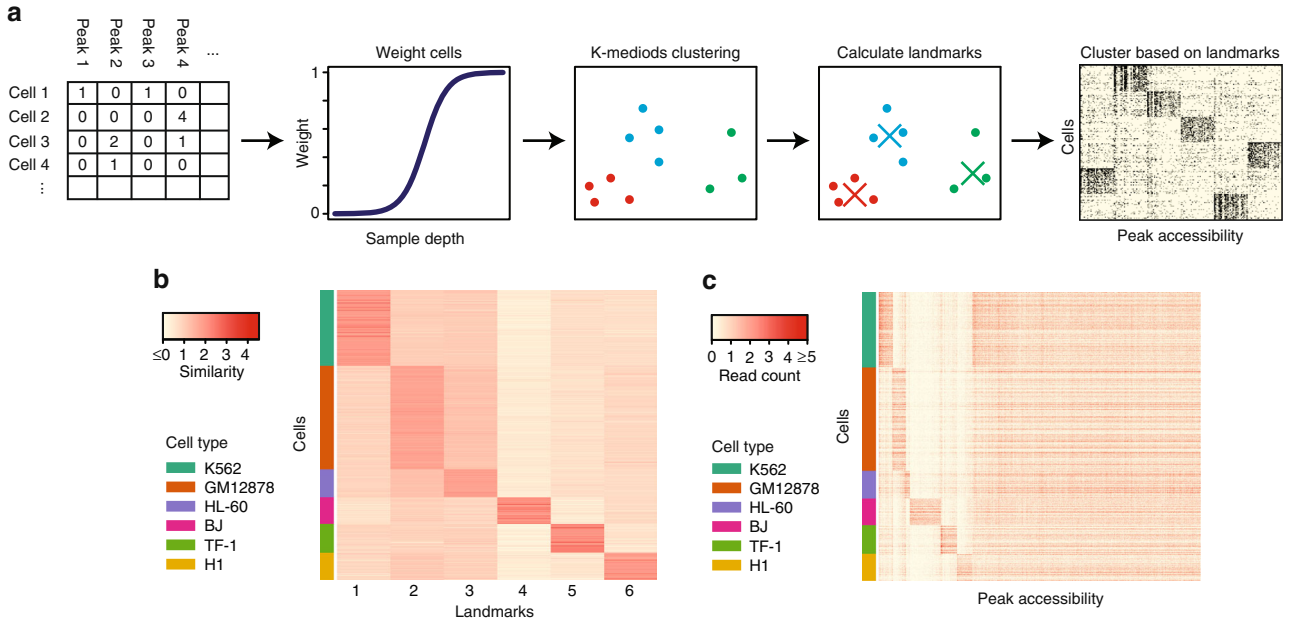

Fig. 1 The *scABC* framework for unsupervised clustering of scATAC-seq data. **a** Overview of *scABC* pipeline. *scABC* constructs a matrix of read counts over peaks, then weights cells by sample depth and applies a weighted *K*-medoids clustering. The clustering defines a set of *K* landmarks, which 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

(Supplementary Note, Supplementary Figs. 1 and 2, and Supplementary Table 1). We then applied *scABC* to this data and determined that there are $K = 6$ clusters using a modified gap statistic (Supplementary Note, Supplementary Fig. 3). We found 6 well separated landmarks with each cell highly correlated with only one landmark (Fig. 1b). The clustering was highly specific with only 4 out of 966 cells misclassified, an error rate of ≈0.4% (Supplementary Table 2).

Three major issues are associated with the in silico mixture that do not appear in natural mixtures. First, the constructed mixture is inherently biased by batch effects since each cell type must be processed separately. To assess the effect of such bias in our method, we noted that the GM12878 cell line was processed in four separate batches, each with the same treatment. We applied *scABC* on the combined four batches of GM12878 cells and the results suggested that there is only a single cluster (Supplementary Fig. 3). To further study batch effects, we intentionally set the number of clusters equal to the number of batches. We found that 99% of the cells were associated with two clusters that have similar landmarks and are not dominated by any batches (Supplementary Fig. 4 and Supplementary Tables 3 and 4). We will investigate these two clusters in a later section but these results indicate that *scABC* is robust to batch effects.

The second major issue is that each distinct cell line makes up at least 9% of the in silico mixture. We tested how the representation of each sub-population affects discovery by reducing the representation of each cell line in the mixture. We found that some well separated sub-populations, such as BJ and TF1, can be distinguished at 1% of the total population, while other sub-populations such as K562 and HL-60 (both of which are erythroleukemic) may merge when the representation of one falls below 5% of the total population (Supplementary Fig. 5). The last issue is that the in silico cell lines are fairly distinct, raising the question: to what extent *scABC* can recognize similar cell types. 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[6]. In principle, narrow peaks better capture TF binding sites[7]. To measure the enrichment of TF motifs in individual cells, we applied chromVAR[8] 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[1,2], *SPI1* in HL-60 cells[9], *GATA1*::*TAL1* in K562 cells[10], and *FOS* in BJ cells[11]. 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[12]. 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

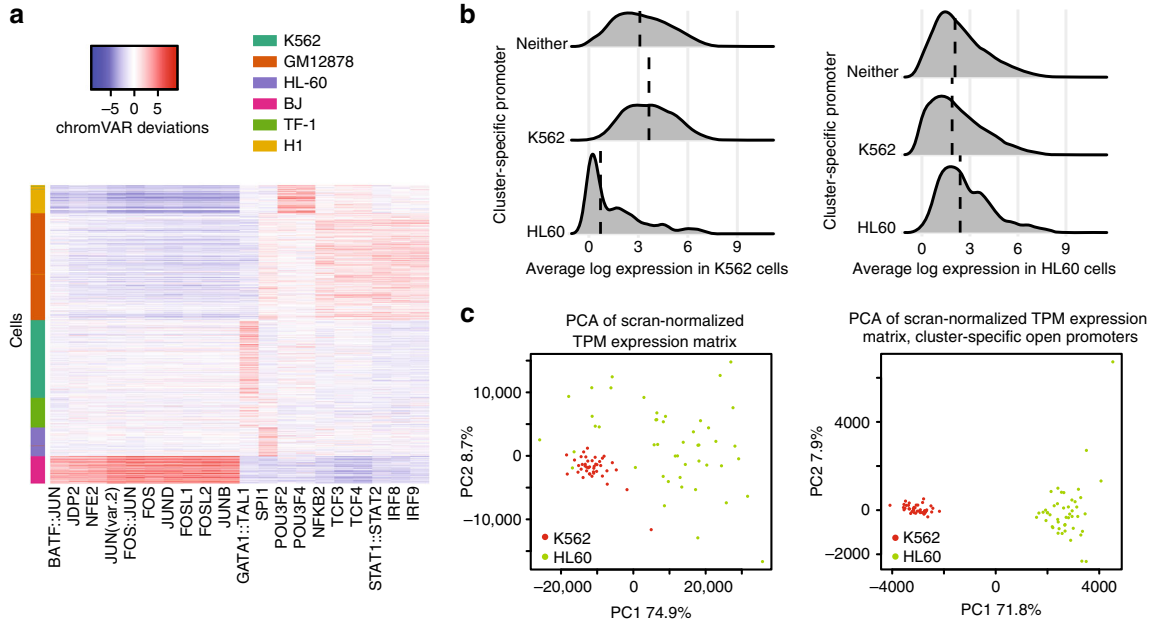

Fig. 2 Cluster specific peaks determined by *scABC* shed light on cell identity. **a** Application of chromVAR to the cluster specific narrow peaks allows for the identification of cluster specific transcription factor binding motifs. chromVAR calculated deviations are shown for the top twenty most variable 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 HL-60 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[13]. 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[14]. To evaluate this hypothesis, we took 42 K562 and 54 HL-60 deeply sequenced scRNA-seq experiments[15] (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[16] 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[17], 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[2]. In 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-κB motifs, such as *NFKB2*, *REL*, and *RELA*, and this may be an indication of transcription factor heterogeneity. The nuclear localization of NF-κB was previously shown to dynamically change and cause temporal variations in transcription factor expression[18], which may explain this heterogeneity. Previous studies[1,2] have also suggested that cellular variability in GM12878 may be driven by NF-κB heterogeneity. These finding 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[19–21]. However, when three-dimensional EBs are treated with RA, previous studies[22,23] 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 *GSX1/2*[24], *LBX1*[25], *LMX1A*[26], *MNX1*[27], *NEUROG2*[28], *NKX6-1/2*[29,30], *UNCX*[31],*VAX1*[32], and POU factors[33–35]. The high activity of *LHX2/9* 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 neurons[36]. In contrast to cluster 1, TFs specific to cluster 2 are essential for visceral endoderm differentiation, such as GATA factors[37], *HNF1A/B*[38,39], and the AP-1 family[40,41] (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[14] (the closest publicly available RNA-seq to day 4). We first used TOMTOM tool[42] to identify TFs with similar motifs to those enriched in cluster 1 and 2 (q-value $\leq 10^{-4}$) 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*, *OLIG1/2*, and *POU3F1* to *POU3F4* 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.[4] sequenced individual monocytes and lymphoid-primed multipotent 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

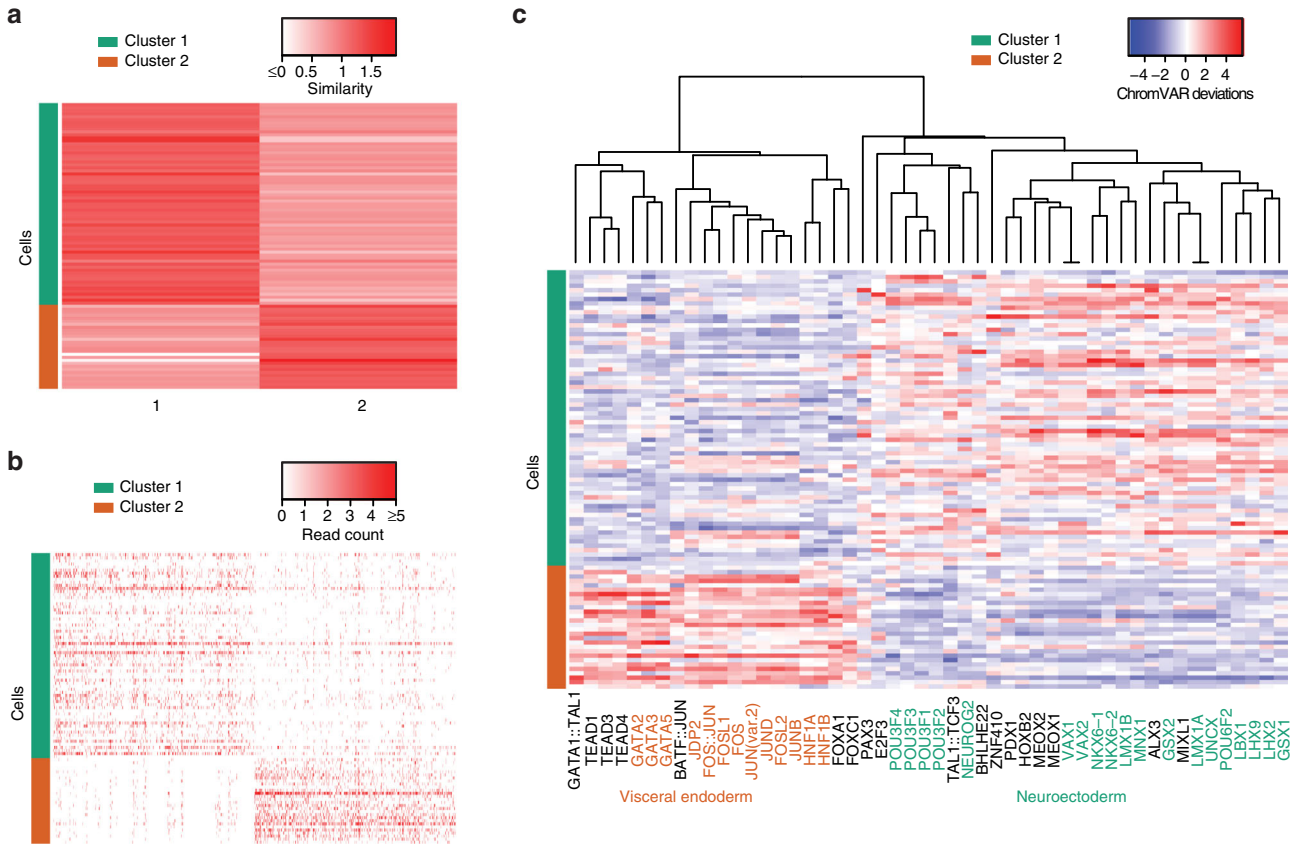

Fig. 3 The application of scABC to a biological cell mixture. **a** 95 scATAC-seq samples were obtained on the day 4 of RA-treated mESC differentiation and classified into two clusters by *scABC*. Here, similarity between cells (rows) and the two detected landmarks (columns) are depicted, with cluster assignments on the left. **b** Heatmap for peak accessibility across cluster specific peaks (columns) and cells (rows). To simplify the presentation for each cluster, we only show the top 500 peaks specific to each cluster, i.e. the smallest scABC p-values (Methods). **c** chromVAR deviations for the top 50 most variable TF motifs (columns) and cells (rows), calculated using cluster specific narrow peaks. Hierarchical cluster analysis of deviations divides motifs into two groups, each specific to just one cluster

clusters, for instance four, the monocyte dominated cluster is stable and well separated from the others but the LMPP is split into two similar clusters (Supplementary Fig. 14 and Supplementary Table 6). Moreover, one cluster contains only LSCs and blasts, which may be an indication of intermediate stages between LMPP and monocyte. Notably, *JUN* and *JUNB* are not enriched in this cluster, and their dysregulation was previously shown to be essential for leukemic stem cell function[43]. In both cases, leukemia cells lie along two major identities on the myeloid progression, represented by monocytes and LMPPs. Our result largely agrees with Corces et al.'s study which was based on separate analysis for each of the 4 cell types[4].

Comparison with previous methods. *scABC* is the first clustering method specifically designed for scATAC-seq. This required us to compare against simpler methods designed for other types of data. We first compared *scABC* against simple *K*-mediods with Spearman dissimilarity measure (without weighting and landmarks) using read counts in peaks and binned counts over long intervals (100 kb, binned using the software csaw[44]), as well as *K*-means on the *log* transcripts per million matrix (with the transcript length equal to the peak length), a common scRNA-seq clustering method.

We applied the above methods to the in silico mixture of six cell lines. To enable a fair comparison, all methods were applied to the cells that pass *scABC* quality control (Supplementary Note). We found that simple *K*-mediods had a slightly higher misclassification rate (1% for *K*-mediods versus 0.4% for *scABC*, Supplementary Tables 2 and 7) while *K*-means on the *log* TPM matrix performed worse (17.3%, Supplementary Table 8) and was not able to separate GM12878 from H1, two distinct cell lines. Clustering 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[45] (a clustering method designed specifically for scRNA-seq) and a community structure clustering method based on the infomap algorithm[46,47]. 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*-mediods (with the best performance among the alternative methods) and *scABC*, we downsampled each cell line and found that *scABC* is able to

identify smaller subpopulations (Supplementary Fig. 15). We next applied *K*-mediods 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, SCDE[48] (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 cannot be separated into subpopulations, such as when 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 embedding[49] or *k*-mer analysis[8,50] 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, named sample depth (Supplementary Note). The weight for cell i is defined as

$$w_i = \frac{1}{1+\exp\{-(h_i-c)/(c\lambda)\}},$$

where c and λ 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_k the medoid for cluster k, i.e. a cell used as the cluster center. The clustering assignment is given by the solution to

$$\underset{C,i_k,k=1,\cdots,K}{\text{minimize}} \sum_{k=1}^{K}\sum_{C(i)=k} w_i d\left(Y_i, Y_{i_k}\right),$$

where $d(\cdot)$ in general represents the dissimilarly 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) algorithm[51] as implemented in the R package *WeightedCluster*[5].

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 statistic[52] with a few modifications to better capture the data structure of single cell experiments, particularly sparsity and cell heterogeneity (Supplementary Fig. 3, 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_{ik} the cluster membership for cell i with $x_{ik} = 1$ if cell i belongs to cluster k and $x_{ik} = 0$ otherwise. We assume that y_{ri} follows a Poisson distribution with mean μ_{ri}.

$$\begin{aligned} y_{ri} &\sim \text{Poisson}(\mu_{ri}),\\ \mu_{ri} &= h_i q_{ri},\\ \log q_{ri} &= \beta_0 + \sum_{k=1}^{K} x_{ik}\beta_{rk},\\ \beta_{rk} &\sim \mathcal{N}(0,\sigma_k^2), \text{for} k = 1,\cdots,K. \end{aligned}$$

The coefficient β_0 is the intercept and the coefficients β_{rk} exhibits 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 β_{rk}, and stabilizes the noisy estimate when the read counts are low[53]. To obtain a robust empirical prior estimate $\hat{\sigma}_k$, we adopt the quantile matching method proposed in DESeq2[53]. In particular, we first fit a model without the intercept β_0 and without the normal prior to attain the maximum likelihood estimate (MLE) $\boldsymbol{\beta}^{mle}$. Let $\bar{\beta}_r^{mle} = \sum_{k=1}^{K}\beta_{rk}^{mle}/K$; let $\bar{\beta}_\cdot^{mle}$ denote the vector $\left(\bar{\beta}_r^{mle}\right)_{r=1,\cdots,R}$; let $\beta_{\cdot k}^{mle}$ indicate the vector $\left(\beta_{rk}^{mle}\right)_{r=1,\cdots,R}$; let $\Phi\left(\cdot;\left|\beta_{\cdot k}^{mle}-\bar{\beta}_\cdot^{mle}\right|\right)$ be the empirical cdf of $\left|\beta_{\cdot k}^{mle}-\bar{\beta}_\cdot^{mle}\right|$, with $\Phi^{-1}\left(\alpha;\left|\beta_{\cdot k}^{mle}-\bar{\beta}_\cdot^{mle}\right|\right)$ equal to the $1-\alpha$ quantile of the empirical cdf; and let z_α be the $1-\alpha$ standard normal quantile. The empirical prior estimate for the standard deviation is calculated as

$$\hat{\sigma}_k = \frac{\Phi^{-1}\left(q;\left|\beta_{\cdot k}^{mle}-\bar{\beta}_\cdot^{mle}\right|\right)}{z_{q/2}}, \text{for} k = 1,\cdots,K.$$

We set $q = 0.05$ in practice. Details for computing $\boldsymbol{\beta}^{mle}$ are described in the Supplementary Note.

Hypothesis testing: Suppose $\Gamma_{-k} = \{1, \ldots, K\}_{-k}$ represent the set $\{1, \ldots, K\}$ except for the kth element. To test whether peak r is specific to cluster k, we consider

$$\begin{aligned} &\text{The null hypothesis } H_0 : \beta_{rk} \le \beta_{rk'}, \text{ for some } k' \in \Gamma_{-k}\\ &\text{The alternative hypothesis } H_1 : \beta_{rk} > \beta_{rk'}, \text{ for all } k' \in \Gamma_{-k}. \end{aligned}$$

Following the intersection-union test[54], the null hypothesis can be broken into $K-1$ simpler null hypotheses $H_{0k'} : \beta_{rk} \le \beta_{rk'}$, with $k' \in \Gamma_{-k}$. For each null hypothesis, the Wald test statistics is $\left(\hat{\beta}_{rk}-\hat{\beta}_{rk'}\right)/SE\left(\beta_{rk}-\beta_{rk'}\right)$, where $\widehat{\boldsymbol{\beta}}$ is the maximum a posteriori (MAP) estimate for $\boldsymbol{\beta}$ 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 α is

$$\frac{\hat{\beta}_{rk}-\hat{\beta}_{rk'}}{\text{SE}\left(\beta_{rk}-\beta_{rk'}\right)} > z_\alpha,$$

and the rejection region for H_0 with level α is

$$\inf_{k'\in\Gamma_{-k}} \frac{\hat{\beta}_{rk}-\hat{\beta}_{rk'}}{\text{SE}\left(\beta_{rk}-\beta_{rk'}\right)} > z_\alpha.$$

Details for computing $\widehat{\boldsymbol{\beta}}$ and the standard errors are illustrated in the Supplementary Note. We finally compute the p-value for H_0 as $\max\{p_{k'}, k' \in \Gamma_{-k}\}$, with $p_{k'}$ 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 U/mL Penicillin-Streptomycin with the addition of 1,000 U/mL of LIF (ESGRO, Millipore).

Cell differentiation: mESCs were differentiated using the hanging drop method[55]. Trypsinized cells were suspended in differentiation medium (mESC medium without LIF) to a concentration of 37,500 cells/ml. 20 μl drops (750 cells) were then placed on the lid of a bacterial plate and the lid was upside down. After 48 h incubation, EBs formed at the bottom of the drops were collected and grown in the well of a 6-well ultra-low attachment plate with fresh differentiation medium containing 0.5 μM RA for 4 days, with the medium being changed daily.

scATAC-seq: We followed the scATAC-seq protocol published by Buenrostro et al.[1] with the following modifications. The EBs were first treated with StemPro® Accutase Cell Dissociation Reagent (Thermo Fisher) at 37 °C for 10–15 min, followed by vigorous pipetting for another 10 min. The cells were passed through 20 μM cell strainer (pluriSelect) to remove un-dissociated EBs. Before loading, the cells were washed three times in C1 DNA Seq Cell Wash Buffer (Fluidigm). In total 9 μL cells 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 the 10–17 μM Fluidigm IFC. Single cells were captured using the "ATACseq: Cell Load and Stain (1861x/1862x/1863x)" scripts. After cell capture, IFC was transferred to a Leica CTR 6000 microscope for imaging, followed by Tn5 transposition and primary 8 cycles of PCR using the "ATACseq: Sample Prep (1861x/1862x/1863x)" scripts. The entire volume (3.5–5 μL) of the amplified transposed DNA was transferred to a 96-wll 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 μM 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 MinElute 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 bps).

Code availability. The *scABC* package is available as an open source R package at https://github.com/timydaley/scABC.

Data 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 typologies of trajectories in the social sciences with R. *LIVES Work. Pap.* **24**, 1–34 (2013).
6. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
7. Ji, H. et al. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.* **26**, 1293–1300 (2008).
8. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
9. 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).
10. Gerstein, M. B. et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
11. Saretzki, G., Feng, J., Zglinicki, T.v. & Villeponteau, B. Similar gene expression pattern in senescent and hyperoxic-treated fibroblasts. *J. Gerontol. A Biol. Sci. Med. Sci.* **53**, B438–B442 (1998).
12. Wang, Z., Oron, E., Nelson, B., Razis, S. & Ivanova, N. Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. *Cell Stem. Cell* **10**, 440–454 (2012).
13. Salmon-Divon, M., Dvinge, H., Tammoja, K. & Bertone, P. PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinf.* **11**, 415 (2010).
14. Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W. H. Modeling gene regulation from paired expression and chromatin accessibility data. *Proc. Natl Acad. Sci. USA* **114**, E4914–E4923 (2017).
15. Pollen, A. A. et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32**, 1053–1058 (2014).
16. Lizio, M. et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16**, 22 (2015).
17. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
18. Tay, S. et al. Single-cell NF-κB dynamics reveal digital activation and analogue information processing. *Nature* **466**, 267–271 (2010).
19. Bain, G., Ray, W. J., Yao, M. & Gottlieb, D. I. Retinoic acid promotes neural and represses mesodermal gene expression in mouse embryonic stem cells in culture. *Biochem. Biophys. Res. Commun.* **223**, 691–694 (1996).
20. Fraichard, A. et al. In vitro differentiation of embryonic stem cells into glial cells and functional neurons. *J. Cell Sci.* **108**, 3181–3188 (1995).
21. Bain, G., Kitchens, D., Yao, M., Huettner, J. E. & Gottlieb, D. I. Embryonic stem cells express neuronal properties in vitro. *Dev. Biol.* **168**, 342–357 (1995).
22. Santos, J. et al. Differences in the epigenetic and reprogramming properties of pluripotent and extra-embryonic stem cells implicate chromatin remodelling as an important early event in the developing mouse embryo. *Epigenet. Chromatin* **3**, 1 (2010).
23. Cai, C. & Grabel, L. Directing the differentiation of embryonic stem cells to neural stem cells. *Dev. Dyn.* **236**, 3255–3266 (2007).
24. Pei, Z. et al. Homeobox genes Gsx1 and Gsx2 differentially regulate telencephalic progenitor maturation. *Proc. Natl Acad. Sci. USA* **108**, 1675–1680 (2011).
25. Müller, T. et al. The homeodomain factor lbx1 distinguishes two major programs of neuronal differentiation in the dorsal spinal cord. *Neuron* **34**, 551–562 (2002).
26. Millonig, J. H., Millen, K. J. & Hatten, M. E. The mouse Dreher gene Lmx1a controls formation of the roof plate in the vertebrate CNS. *Nature* **403**, 764 (2000).
27. Wichterle, H., Lieberam, I., Porter, J. A. & Jessell, T. M. Directed differentiation of embryonic stem cells into motor neurons. *Cell* **110**, 385–397 (2002).
28. Ribes, V. et al. Combinatorial signalling controls Neurogenin2 expression at the onset of spinal neurogenesis. *Dev. Biol.* **321**, 470–481 (2008).
29. Li, X. et al. Specification of motoneurons from human embryonic stem cells. *Nat. Biotechnol.* **23**, 215–221 (2005).
30. Okada, Y., Shimazaki, T., Sobue, G. & Okano, H. Retinoic-acid-concentration-dependent acquisition of neural cell identity during in vitro differentiation of mouse embryonic stem cells. *Dev. Biol.* **275**, 124–142 (2004).
31. Asbreuk, C., Van Doorninck, J., Mansouri, A., Smidt, M. & Burbach, J. Neurohypophysial dysmorphogenesis in mice lacking the homeobox gene Uncx4.1. *J. Mol. Endocrinol.* **36**, 65–71 (2006).
32. Hallonet, M., Hollemann, T., Pieler, T. & Gruss, P. Vax1, a novel homeobox-containing gene, directs development of the basal forebrain and visual system. *Genes Dev.* **13**, 3106–3114 (1999).
33. Zhu, Q. et al. The transcription factor Pou3f1 promotes neural fate commitment via activation of neural lineage genes and inhibition of external signaling pathways. *eLife* **3**, e02224 (2014).
34. Pang, Z. P. et al. Induction of human neuronal cells by defined transcription factors. *Nature* **476**, 220–223 (2011).
35. Sugitani, Y. et al. Brn-1 and Brn-2 share crucial roles in the production and positioning of mouse neocortical neurons. *Genes Dev.* **16**, 1760–1765 (2002).
36. Sharma, K. et al. LIM homeodomain factors Lhx3 and Lhx4 assign subtype identities for motor neurons. *Cell* **95**, 817–828 (1998).
37. Fujikura, J. et al. Differentiation of embryonic stem cells is induced by GATA factors. *Genes Dev.* **16**, 784–789 (2002).
38. Coffinier, C., Thépot, D., Babinet, C., Yaniv, M. & Barra, J. Essential role for the homeoprotein vHNF1/HNF1beta in visceral endoderm differentiation. *Development* **126**, 4785–4794 (1999).
39. Cereghini, S., Ott, M., Power, S. & Maury, M. Expression patterns of vHNF1 and HNF1 homeoproteins in early postimplantation embryos suggest distinct and sequential developmental roles. *Development* **116**, 783–797 (1992).
40. Xanthoudakis, S., Smeyne, R. J., Wallace, J. D. & Curran, T. The redox/DNA repair protein, Ref-1, is essential for early embryonic development in mice. *Proc. Natl Acad. Sci. USA* **93**, 8919–8923 (1996).

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**, 465–471 (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**, e8694 (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.