Unsupervised clustering and epigenetic classification of single cells

Mahdi Zamanighomi[1,*], Zhixiang Lin[1,*], Timothy Daley[1,2,*], Alicia Schep[3,4], William J Greenleaf[3,4,5] & Wing Hung Wong[1,6,†]

[1]Department of Statistics, Stanford University, Stanford, CA 94305, USA
[2]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA
[3]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA
[4]Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA 94305
[5]Department of Applied Physics, Stanford University, Stanford, CA 94305, USA
[6]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA

[*]Authors contributed equally to this work
[†]Correspondence should be addressed to: whwong@stanford.edu

Abstract

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 heterogenous 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[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. For instance, 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 subpopulations, 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, named *scABC* (single cell Accessibility Based Clustering) and outlined in Fig. 1**a**. 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 a merger of all individual cells without the need for additional experiments). We apply our method to publicly available scATAC-seq data[1,2,4] 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 have functional meaning and are capable of determining cellular identity. In particular, we show that the cluster specific accessible regions are enriched for transcription factor motifs known to be specific to each subpopulation and also, through association with scRNA-seq data, leads to the identification of subpopulation specific gene expression.

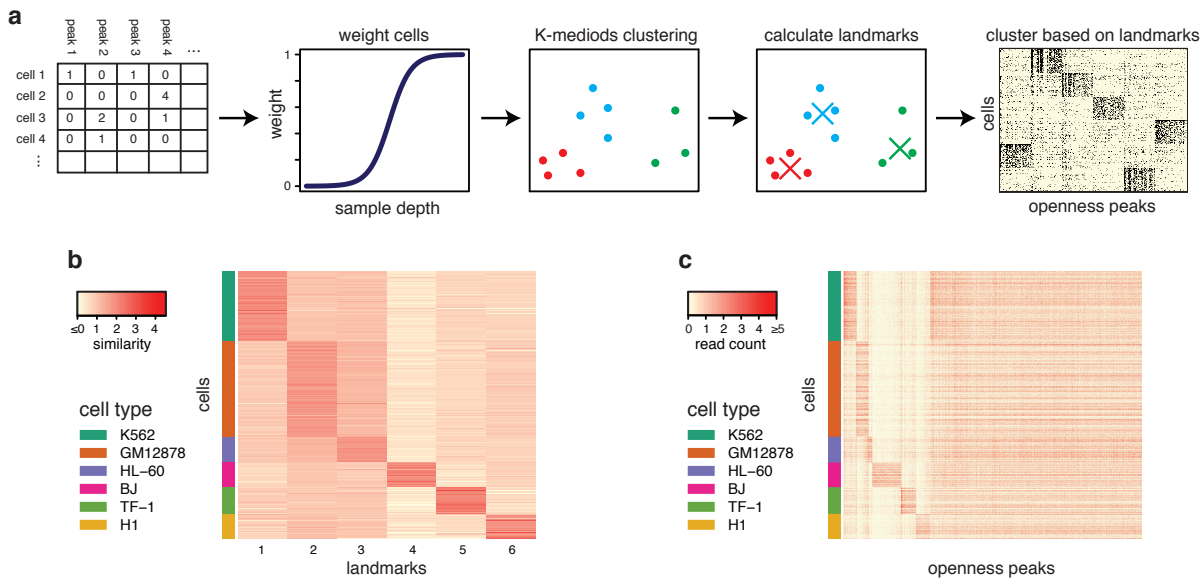

Figure 1: **The *scABC* framework for unsupervised clustering of scATAC-seq data.** (**a**) Overview of *scABC* pipeline. *scABC* constructs a matrix of read counts over peaks, then weights cells by sample depth and applies a weighted K-mediods clustering. The clustering defines a set of K landmarks, which are then used to reassign cells to clusters. (**b**) Assignment of cells to landmarks by Spearman correlation, where each cell highly correlated with just one landmark. Here, similarity is defined as the Spearman correlations of each cell and landmarks, normalized by the absolute value of their mean. The normalization is used to simplify the comparison of landmarks and is not considered in our analysis. (**c**) Openness peaks across all cells. The vast majority of peaks tend to be either common or cluster specific, allowing us to define cluster specific peaks.

First, we briefly describe our algorithm and the intuition behind it. 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 transform of) the number of distinct reads within peak backgrounds and then applies a weighted K-mediods clustering[5] to partition the cells into distinct groups (see Methods for details). To prevent bias from jackpots and highly over-represented regions, *scABC* uses the ranked peaks in each cell to perform the clustering rather than the raw counts. Next, the clustering results are used to calculate "landmarks" for each identified cluster. These landmarks represent prototypical cells from each cluster and we reassign cells to cluster based on the highest Spearman correlation to the respective landmarks. With the cluster assignments we finally test whether each accessible region is specific to a particular cluster, using an empirical Bayes regression based hypothesis testing procedure to obtain cluster specific peaks (Methods)

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] (details in Supplementary Notes and Table S1). We then applied *scABC* to this data and determined that there are $K = 6$ clusters (see Supplementary Notes for the determination of K, Supplementary Fig. S1). We found 6 well separated landmarks with each cell highly correlated with only one landmark (Fig. 1**b**). The clustering was highly specific and only 4 out of 966 cells were misclassified, an error rate of $\approx 0.4\%$ (Supplementary Table S2). Furthermore, we observed that the majority of peaks fall into one of two categories: peaks specific to each cluster and peaks common to all cells (Fig. 1**c**).

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 cells were processed and sequenced in several 96-well plates[1]. To assess the effect of such bias in our method, we noted that the GM12878 cell line was processed in 4 batches, the highest number among the 6 cell lines. We ran *scABC* on the combined 4 batches of GM12878 cells

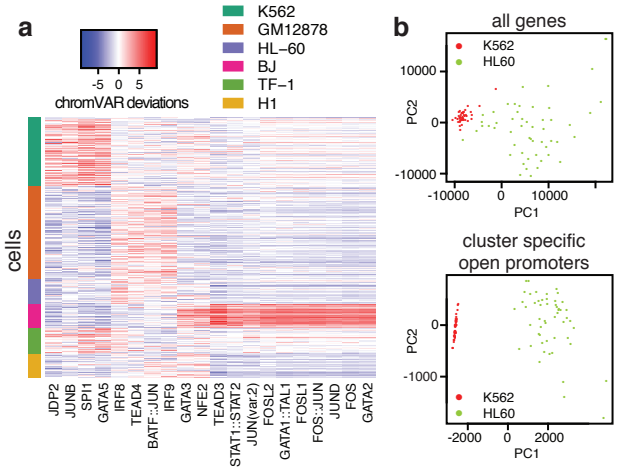

Figure 2: **Cluster specific peaks, attained by *scABC*, shed light on cell identity.** (**a**) Application of chromVAR to the cluster specific narrow peaks allows for the identification of cluster specific transcription factor binding motifs. chromVAR calculated deviations are shown for the top twenty most variable transcription factor binding motifs. (**b**) 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 (bottom) and compared to PCA of all genes (top).

and obtained $K = 1$ cluster (Supplementary Fig. S1), however, we set the number of clusters equal to the number of batches to further analyze batch effects. 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. S2 and Tables S3-S4). Although there was no indication of clear batch effect in the *scABC* cluster assignment, we hypothesized that small variations between the two clusters may suggest the heterogeneity of GM12878 cell line. We then applied *scABC*, with $K = 2$, to GM12878 cells and observed two somewhat separated clusters. We also observed that one cluster is notably enriched for NF-κB motifs (Supplementary Fig. S3, motif analysis is described in the next few paragraphs), consistent with previous findings[1,6]. To further investigate the batch effect, we applied *scABC* on a experimental mixture of GM12878 and HEK293T cells as well as GM12878 and HL-60[2], where cells were processed in a single batch for each mixture. In both cases, clear separation between the two cell lines were achieved (Supplementary Fig. S4 and S5), which, due to the experimental design, can not be explained by batch effects.

The second major issue is that each distinct cell line is present in 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 H1, can be distinguished at 2% of the total population, while similar sub-populations, such as GM12878 and HL-60, will merge when the representation of one falls below 5% (Supplementary Fig. S6). The last issue is that the *in silico* cell lines are fairly distinct, raising the question: to what extend *scABC* can recognize similar cell types. Due to the lack of public scATAC-seq data for similar cell lines, we designed a test to evaluate *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. S7).

We next investigated whether the cluster specific peaks, obtained by *scABC*, are able to define cell identity. These peaks contain both narrow and broad regions, as defined by MACS2[7]. In principle, narrow peaks better reflect TF binding sites[8]. We applied chromVAR[9] to *scABC* defined cluster specific narrow peaks for the discovery of enriched cluster specific TF motifs. We found that the most active TFs are specific to one or two clusters, identifying active TFs in each cell type (Fig. 2**a**). For instance, BATF-JUN complex is quite active in GM12878 cells and the transcription factors JUN and FOS in BJ cells, in agreement with previous analyses[10,11]. Note that BJ specific TFs are quite

enriched, but that is expected since the BJ cluster is better separated from the other clusters and has a higher number of cluster specific peaks (Fig. 1**b** and **c**). We also applied chromVAR to the full set of narrow peaks and found comparable results (Supplementary Fig. S8), 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[12]. We hypothesized that cluster specific broad peaks overlapping gene promoters have functional significance and can help distinguish genes specific to a particular cell type. To evaluate our hypothesis, we took 42 K562 and 54 HL-60 deeply sequenced scRNA-seq experiments[13] and clustered the corresponding gene expression data (in transcripts per million) using both all genes and only those genes with cell type specific open promoters in K562 and HL-60 cells (as shown in Fig. 2**b**). After normalization for dropout-associated batch effects[14], clustering based on all genes did not clearly separate the two cell types in the first two principal components (consistent with the results of Hicks *et al.*[14]). When genes associated with cell-type specific open promoters were employed, the separation became extremely obvious. This indicates two important results: i) cluster specific broad peaks shed light on functional significance outside of motif enrichment and ii) our approach enables joint analysis of single cell epigenetic and single cell transcriptional data to improve functional analysis of gene regulation.

In addition to the *in silico* cell line mixture, we examined *scABC*'s 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. S9). We applied *scABC* followed by chromVAR to the combined mixture of the 390 cells that passed quality control (Supplementary Notes). 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 Fig. S1, Fig. S10, and Table S5, details in Supplementary Notes). When using more 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. S11 and Table S6). Moreover, one cluster contains only LSCs and blasts, which might be an indication of intermediate stages between LMPP and monocyte. 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].

In summary, we developed a tool for the unsupervised analysis of single cell epigenetic data, yielding promising results on scATAC-seq of complex mixtures. In combination with other tools such as chromVAR, *scABC* enables deeper understanding of regulatory heterogeneity that were previously hard to ascertain. *scABC* is available as an open source R package at `https://github.com/timydaley/scABC`.

Acknowledgments

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

Author Contribution

M.Z, Z.L, T.D, and W.H.W convinced the project. M.Z, Z.L, T.D, and A.N.S performed the analysis with inputs from W.J.G and W.H.W. M.Z, Z.L, and T.D prepared *scABC* R package. All authors wrote the manuscript.

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 (details in Supplementary Notes). 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. 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

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

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 Mediods (PAM) algorithm[15] as implemented in the R package $WeightedCluster$[5].
Landmarks: We sum the reads across the cells within a cluster and select the 2,000 peaks with the highest read counts to obtain the landmark for each cluster identified in the previous step.
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.

Incorporating the weighting scheme for diverse sample depths slightly improves the separation of cell lines in the *in silico* mixture cell population (Supplementary Tables S2 and S7). In addition, the weighted K-medoids distinguishes smaller cell sub-populations compared to the unweighted K-medoids algorithm (Supplementary Fig. S6).

The weighted K-medoids algorithm requires the number of clusters K in advance. We determine K through the gap statistic[16] with a few modifications to better capture the data structure of single cell experiments, particularly sparsity and cell heterogeneity (Supplementary Fig. S1, details in Supplementary Notes).

For the tuning parameters, we set $c = \min(8,\ \text{median}(h))$ and $\lambda = 0.1$ in practice. Notably, the performance of the clustering is robust to a wide range of $\{c, \lambda\}$ (Supplementary Table S8).

Identification of cluster specific peaks

We formulate this problem in a hypothesis testing framework to find peaks that tend to be more open in one cluster compared to all other clusters. We perform the hypothesis testing on the union of narrow and broad peaks, but, the procedure is applicable to any set of peaks. Here, 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}.

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

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[17]. To obtain a robust empirical prior estimate $\hat{\sigma}_k$, we adopt the quantile matching method proposed in DESeq2[17]. 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 $\bar{\beta}_{r\cdot}^{mle} = \sum_{k=1}^{K} \beta_{rk}^{mle}/K$; let $\bar{\beta}_{\cdot\cdot}^{mle}$ denote the vector $(\bar{\beta}_{r\cdot}^{mle})_{r=1,\cdots,R}$; let $\beta_{\cdot k}^{mle}$ indicate the vector $(\beta_{rk}^{mle})_{r=1,\cdots,R}$; let $\Phi_{|\beta_{\cdot k}^{mle}-\bar{\beta}_{\cdot\cdot}^{mle}|}$ be the empirical cdf of $|\beta_{\cdot k}^{mle} - \bar{\beta}_{\cdot\cdot}^{mle}|$, with $\Phi^{-1}_{|\beta_{\cdot k}^{mle}-\bar{\beta}^{mle}|}(\alpha)$ 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

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

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

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

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

Following the intersection-union test[18], the null hypothesis can be broken into $K-1$ simpler null hypotheses $H_{0k'} : \beta_{rk} \leq \beta_{rk'}$, with $k' \in \Gamma_{-k}$. For each null hypothesis, the Wald test statistics is $(\hat{\beta}_{rk} - \hat{\beta}_{rk'})/SE(\beta_{rk} - \beta_{rk'})$, where $\hat{\boldsymbol{\beta}}$ 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

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

and the rejection region for H_0 with level α is

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

Details for computing $\hat{\boldsymbol{\beta}}$ and the standard errors are illustrated in the Supplementary Notes. 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 $H_{0k'}$.

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. *Nature biotechnology* **33**, 1165–1172 (2015).

4. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature genetics* **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 Working Papers* **24**, 1–34 (2013).

6. Tay, S. *et al.* Single-cell NF-κB dynamics reveal digital activation and analogue information processing. *Nature* **466**, 267–271 (2010).

7. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nature protocols* **7**, 1728–1740 (2012).

8. Ji, H. *et al.* An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature biotechnology* **26**, 1293–1300 (2008).

9. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: Inferring transcription factor variation from single-cell epigenomic data. *bioRxiv* 110346 (2017).

10. Arvey, A., Agius, P., Noble, W. S. & Leslie, C. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome research* **22**, 1723–1734 (2012).

11. Kovary, K. & Bravo, R. The jun and fos protein families are both required for cell cycle progression in fibroblasts. *Molecular and cellular biology* **11**, 4466–4472 (1991).

12. Salmon-Divon, M., Dvinge, H., Tammoja, K. & Bertone, P. PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC bioinformatics* **11**, 415 (2010).

13. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology* **32**, 1053–1058 (2014).

14. Hicks, S. C., Teng, M. & Irizarry, R. A. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-seq data. *bioRxiv* 025528 (2015).

15. Rousseeuw, P. J. & Kaufman, L. *Finding Groups in Data* (Wiley Online Library, Hoboken, New Jersey, USA, 1990).

16. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**, 411–423 (2001).

17. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).

18. Casella, G. & Berger, R. L. *Statistical inference* (Duxbury, Pacific Grove, California, USA, 2002).