Learning relationships between histone modifications in single cells

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent advances have enabled mapping of histone marks in single cells, but most methods are constrained to profile only one histone mark per cell. Here we present an integrated statistical and experimental framework, scChIX (single-cell chromatin immunocleavage and unmixing), to map multiple marks in single cells. scChIX multiplexes two marks together in single cells, then computationally deconvolves the signal using training data from respective histone mark profiles. This framework learns the cell type-specific correlation structure between histone marks, and therefore does not require *a priori* assumptions of their genomic distributions. Applying scChIX to two active marks during *in vitro* macrophage differentiation, we find H3K4me1 dynamics preceding H3K36me3. Modeling these dynamics enables integrated analysis of chromatin velocity during differentiation. Overall, scChIX reveals unique biological insights by leveraging multimodal analysis between histone modifications in single cells.

1 Introduction

Mapping histone modifications in single cells is still in its infancy, and has the potential to uncover how the chromatin is read to govern gene expression. Currently, most experimental techniques (e.g. scChIC-seq [1, 2, 3], CUT&RUN [4], and CUT&TAG [5, 6, 7, 8, 9, 10]) that map single-cell histone modifications are limited to only one histone modification per single cell.

We present an integrated experimental and computational framework for multiplexing histone modifications in single cells. To profile two histone modifications in single cells (Figure 1a), we first generate three genome-wide sortChIC [1] datasets: two datasets by incubating cells with one of the two histone modification antibodies separately (single-incubated, Figure 1b), and the third by incubating cells with both histone modification antibodies together (double-incubated, Figure 1b). We then use our two single-incubated datasets as training data to generate the possible pairs of genome-wide histone modification profiles that, when added together, fit to a single-cell profile from the double-incubated dataset (Figure 1c). For each double-incubated cell, we then deconvolve the multiplexed data by probabilistically assigning each fragment back to their respective histone modification.

scChIX links single-cell maps different histone modifications, revealing relationships between histone modifications in single cells. In these linked maps, information derived from one chromatin state, such as cell types, histone mark levels, and pseudotimes, can transfer to another chromatin state (Figure 1d), unlocking joint analysis of multiple histone modifications in single cells. We first validate scChIX using simulation, then apply scChIX to macrophage *in vitro* differentiation to reveal coordinated dynamics between H3K4me1 and H3K36me3.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

2 Methods

Unmixing scChIX signal

Single-cell epigenomics techniques (e.g. sortChIC, CUT&RUN, CUT&TAG) generate a vector of counts indicating the number of cut fragments that map in each genomic region for each cell. We model the vector of counts from a double-incubated cell $\vec{y}$ as a linear combination of two multinomial distributions, one coming from a cluster c of histone modification 1, parameterized by $\vec{p}_c$, the other from another cluster d of histone modification 2 $\vec{q}_d$. The log-likelihood for a linear combination of two multinomials is:

$$\mathrm{L}_{(c,d)} = \log(P\left(\vec{y}|\vec{p}_c, \vec{q}_d, w\right)) \propto \sum_{g=1}^{G} y_g \log\left(wp_{c,g} + (1-w)\, q_{d,g}\right). \tag{1}$$

$\vec{y}$ is the number of cuts across the genome for a double-incubated cell. $p_{c,g}$ and $q_{d,g}$ are cluster-specific probabilities indicating the likelihood that a cut fragment maps to region g in histone modifications 1 and 2, respectively. w is the mixing fraction of histone modification 1 in the double-incubated cell, which we estimate by maximizing the log-likelihood given $\vec{y}$, $\vec{p}_c$, and $\vec{q}_d$.

Applying single-cell techniques to complex tissues generates data with many clusters. Therefore, given a double-incubated cell, we do not know which pair of clusters (c, d) were combined to generate the observed counts. We therefore calculate the log-likelihood for all possible pairs of clusters learned from the training data and then we select the cluster pair with the highest probability for each cell.

Cluster-specific probabilities $\vec{p}_c$ and $\vec{q}_d$, are learned by applying latent Dirichlet allocation (LDA, with k=30 topics) to the training data (i.e. single-incubated cells), which are also count matrices.

After assigning each cell to the most likely cluster pair $(\hat{c}, \hat{d})$, we assign $y_{i,j}$, the *j*th read mapped to region g in cell i, to histone mark 1 with probability $P_{i,j}$:

$$P_{i,j} = \frac{wp_{\hat{c},g}}{wp_{\hat{d},g} + (1-w)\, q_{\hat{d},g}}. \tag{2}$$

The log-likelihood (Equation 1) can be extended to continuous pseudotime by replacing $\vec{p}_c$ and $\vec{q}_d$ with a probability vector that depends on pseudotime $\vec{p}(t_1)$ and $\vec{q}(t_2)$

Chromatin velocity during macrophage differentiation

We assume that dynamic genomic regions in H3K36me3 can be modeled using a first-order differential equation

$$\frac{dK_{36}\left(t\right)}{dt} = K_4\left(t\right) - \gamma K_{36}\left(t\right). \tag{3}$$

We estimate the time constant γ for each genomic region by fitting an exponential relaxation function across pseudotime

$$K_{36}\left(t\right) = y_0 + A\left(1 - e^{-\gamma t}\right), \tag{4}$$

where y_0 is the signal at $t = 0$ and A is the predicted H3K36me3 levels at steady state. Fitting the γ directly from the pseudotime allows us to leverage signal from both single- and deconvolved cells.

To predict future values of H3K36me3 levels for each cell at each genomic region, we use the Euler method and plug in the estimated γ, H3K4me1 levels at time t, and time step h of 0.02 pseudotime units.

Finally, we project the single- and double-incubated H3K36me3 signal onto the first two principal components and project the predicted future values onto the PCA. We use the velocity grid flow visualization as implemented in velocyto [11] to visualize the velocity vectors on the PCA space.

3 Results

3.1 Simulation study validates that scChIX is applicable to a wide range of overlapping relationships between histone modifications

To test whether scChIX is accurate for histone modification patterns that are mutually exclusive as well as highly overlapping, we apply scChIX to simulated single-cell data with known amounts of overlap to benchmark our method across different overlapping patterns between histone modifications. We simulate single-cell histone modification data by modifying simATAC [12] to generate sparse count data from different overlapping patterns from the same cell (Figure 1e). In these simulations, we provide a ground truth parameter p for each genomic locus and then estimate this parameter using our statistical framework to assess the uncertainty in our inferences. Here, p is the expected fraction of double-incubated reads in a locus that belongs to a reference histone modification (i.e. p=0.5 if locus is exactly overlapping, p=1 or 0 if locus is exactly mutually exclusive). We find that scChIX accurately recovers the different cell types underlying the simulated data, and links the two histone modification landscapes into a joint UMAP (Figure 1f). We find that scChIX can estimate p accurately for all degrees of overlap, with confidence intervals better than $\hat{p} \pm 0.05$ (Figure 1g). Our simulation study confirms that scChIX is accurate in inferring multiple histone modifications in single cells in both mutually exclusive as well as overlapping histone modification patterns.

3.2 scChIX reveals macrophage differentiation dynamics and chromatin velocity

Finally, we apply scChIX to study the dynamic relationships between two active histone modifications, H3K4me1 and H3K36me3, over an *in vitro* differentiation time course. We sort blood progenitors from mouse bone marrow, add macrophage colony-stimulating factor (MCSF), and collect cells over seven days (Figure 2a). We incubate cells with either H3K4me1, H3K36me3, or both H3K4me1 and H3K36me3, then perform scChIX.

Using continuous pseudotime of H3K4me1 and H3K36me3 as our training data (Methods), we infer where along pseudotime each double-incubated cell came from for H3K4me1 and H3K36me3. Plotting the inferred pseudotimes of each mark for each cell uncovers the dynamic relationships between the two marks (Figure 2b). H3K4me1 pseudotime initially progresses while H3K36me3 remains relatively unchanged. As H3K4me1 pseudotime approaches 0.5, H3K36me3 then rapidly progresses towards 1. This sigmoidal-like relationship between H3K4me1 versus H3K36me3 pseudotime progression is consistent with H3K4me1 dynamics occurring before H3K36me3. Finally, we use this inferred pseudotime information to project the deconvolved cells onto the H3K4me1 and H3K36me3 UMAPs. Using the deconvolved cells as anchors, we then link the two histone modification maps together (Figure 2c).

Since we observe that H3K4me1 dynamics occur before H3K36me3, we ask whether we could model the H3K36me3 dynamics as a first-order differential equation analogous to RNA velocity [11]. Since our data comes from a time course, we directly fit the exponential curves for dynamic genes along pseudotime for H3K36me3 (Figure 2d top, Methods), which avoids making steady-state assumptions and leverages information from both single-incubated and deconvolved cells across histone modifications. For each gene, we use the estimated rate constants to predict the H3K36me3 levels for each cell over small pseudotime steps (examples of two genes in Figure 2d). Finally, summarizing the predictions of dynamic genes, we project the high-dimensional velocity vectors onto the first two principal components (Δt = 0.02, Figure 2e). From the chromatin velocity summary, we find that differentiation starts with large changes in H3K36me3 dynamics which then relaxes towards the macrophage state.

In summary, we apply scChIX to two active histone modifications to find dynamic relationships between activation states. We then model these dynamics to infer chromatin velocity during macrophage differentiation.

4 Discussion

Here, we demonstrate that by combining statistical inference with experimental design, we can expand the number of histone modifications that can be profiled in single cells. We show in simulations and real data, that scChIX can infer relationships between histone modifications. We find that scChIX

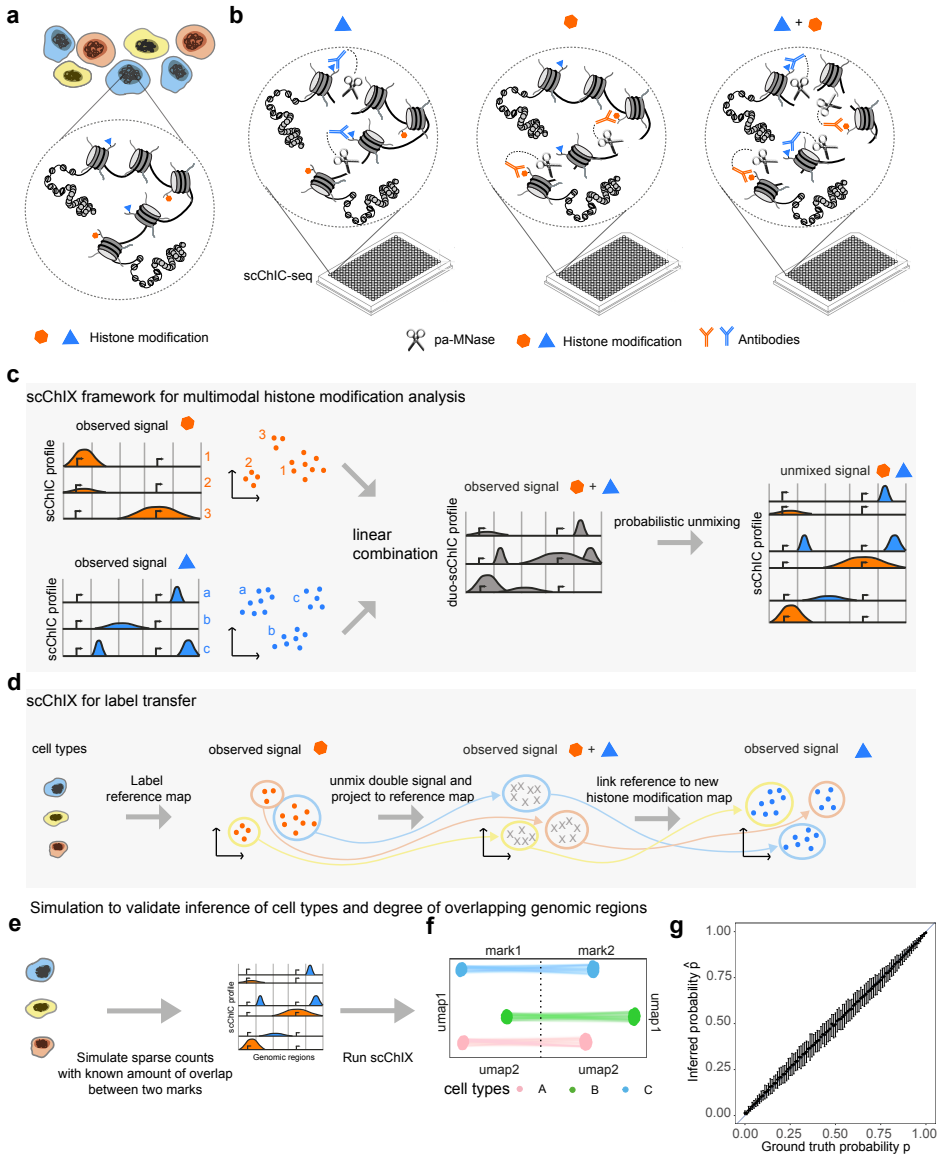

Figure 1: **Overview of the scChIX method.** **(a)** Schematic of a cell with multiple histone modifications. **(b)** Experimental strategy: scChIX uses three sortChIC antibody incubation conditions: two conditions each target a single histone modification only (single-incubated) and the third condition targeting both histone modifications simultaneously (double-incubated). **(c)** Schematic of scChIX for deconvolving multiplexed histone modifications. The two single-incubated sortChIC datasets (orange and blue) are training data. Pairs of distributions can be combined to generate a hypothetical double-incubated cell. For each observed double-incubated cell, we then assign the cell to the most likely pair of cell states, one from each histone modification. We then probabilistically assign each cut fragment to their respective histone modification. **(d)** Schematic for label transfer. **(e)** Schematic of simulation study. **(f)** Joint UMAP of deconvolved cells. **(g)** Comparison of ground truth p versus the estimate $\hat{p}$. p is the expected fraction of double-incubated reads in a genomic locus that belongs to mark 1. $\hat{p}$ is the estimate of the probability. Error bars are 95% confidence intervals.

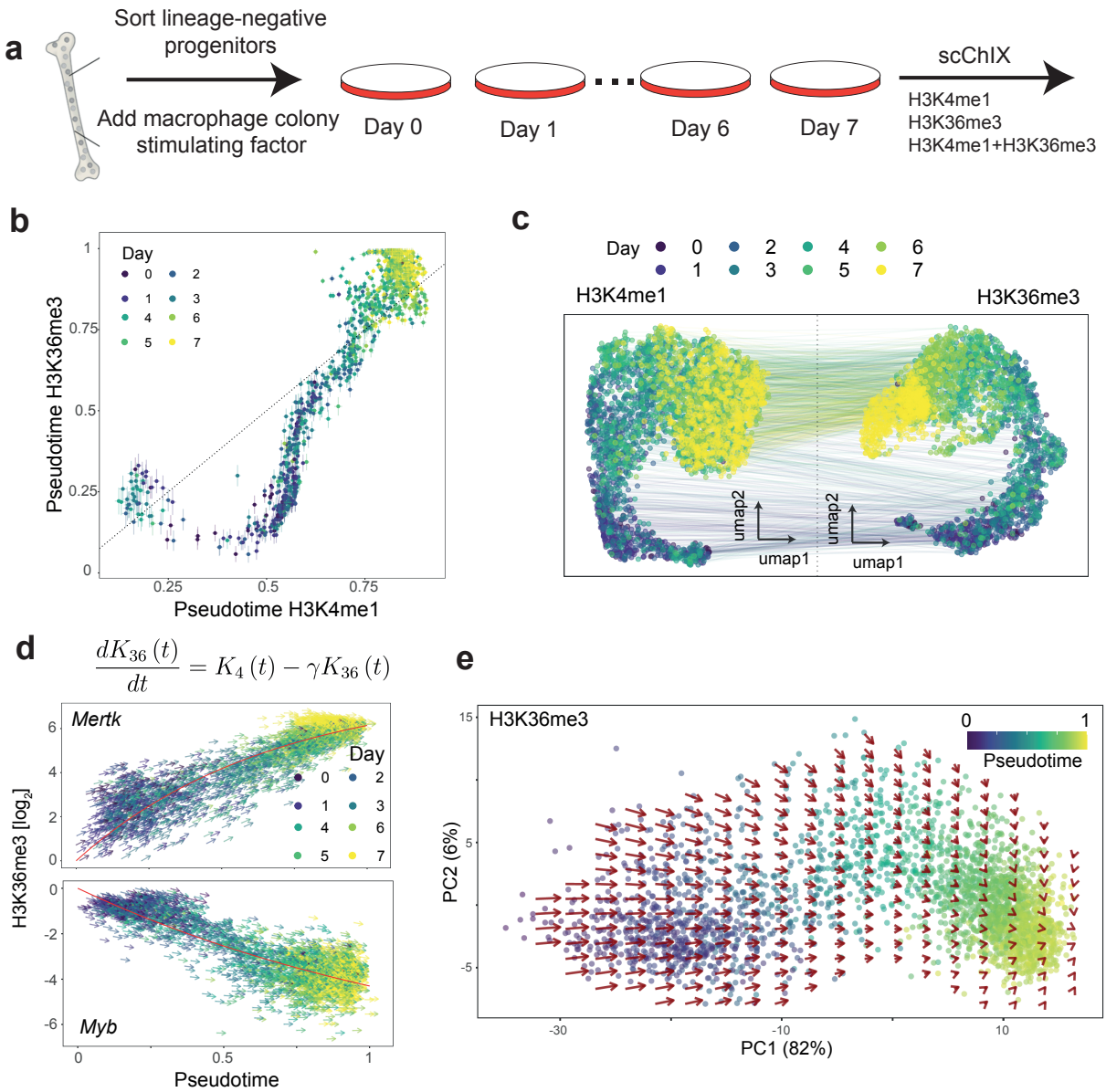

Figure 2: **Applying scChIX to two active marks reveals chromatin velocity during macrophage differentiation. (a)** Schematic of mouse macrophage *in vitro* differentiation time course experiment to study H3K4me1 and H3K36me3 in single cells. **(b)** Estimates of two different pseudotimes from a single cell. Error bars are 95% confidence intervals of the estimates. Each point is a double-incubated cell. **(c)** Joint UMAP of H3K4me1 and H3K36me3 from scChIX, lines connect single cells with multimodal information. **(d)** Chromatin velocity estimates of an upregulated gene (above) and a downregulated gene (below). Red curve is the exponential relaxation fit according to the first-order differentiation equation. **(e)** High-dimensional chromatin velocities of dynamic genes projected onto the principal components one and two.

can reveal biological insights by multimodal analysis that would otherwise be obscured by analyzing each modality separately. Importantly, scChIX is flexible in which histone modifications can be used. The correlation structure between modifications is learned from the model and therefore do not require *a priori* assumptions of specific features of the two modifications. Overall, scChIX unlocks multimodal analysis in antibody-based chromatin profiling and enables joint analysis of distinct histone modifications in single cells.

References

[1] Peter Zeller, Jake Yeung, Buys Anton de Barbanson, Helena Viñas Gaza, Maria Florescu, and Alexander van Oudenaarden. Hierarchical chromatin regulation during blood formation uncovered by single-cell sortChIC. *bioRxiv*, 2021. doi: 10.1101/2021.04.26.440606.

[2] Wai Lim Ku, Kosuke Nakamura, Weiwu Gao, Kairong Cui, Gangqing Hu, Qingsong Tang, Bing Ni, and Keji Zhao. Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. *Nature Methods*, 16(4):323–325, 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0361-7. URL `https://doi.org/10.1038/s41592-019-0361-7`.

[3] Wai Lim Ku, Lixia Pan, Yaqiang Cao, Weiwu Gao, and Keji Zhao. Profiling single-cell histone modifications using indexing chromatin immunocleavage sequencing. *Genome Research*, page gr.260893.120, apr 2021. ISSN 1088-9051. doi: 10.1101/gr.260893.120. URL `http://genome.cshlp.org/lookup/doi/10.1101/gr.260893.120`.

[4] Peter J Skene, Jorja G Henikoff, and Steven Henikoff. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nature Protocols*, 13(5):1006–1019, 2018. ISSN 1750-2799. doi: 10.1038/nprot.2018.015. URL `https://doi.org/10.1038/nprot.2018.015`.

[5] Hatice S. Kaya-Okur, Steven J. Wu, Christine A. Codomo, Erica S. Pledger, Terri D. Bryson, Jorja G. Henikoff, Kami Ahmad, and Steven Henikoff. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, 10(1):1–10, 2019. ISSN 20411723. doi: 10.1038/s41467-019-09982-5.

[6] Qianhao Wang, Haiqing Xiong, Shanshan Ai, Xianhong Yu, Yaxi Liu, Jiejie Zhang, and Aibin He. CoBATCH for high-throughput single-cell epigenomic profiling. *Molecular Cell*, pages 1–11, 2019. ISSN 1097-2765. doi: 10.1101/590661. URL `http://dx.doi.org/10.1101/590661`.

[7] Akihito Harada, Kazumitsu Maehara, Tetsuya Handa, Yasuhiro Arimura, Jumpei Nogami, Yoko Hayashi-Takanaka, Katsuhiko Shirahige, Hitoshi Kurumizaka, Hiroshi Kimura, and Yasuyuki Ohkawa. A chromatin integration labelling method enables epigenomic profiling with lower input. *Nature Cell Biology*, 21(2):287–296, 2019.

[8] Steven J. Wu, Scott N. Furlan, Anca B. Mihalas, Hatice S. Kaya-Okur, Abdullah H. Feroze, Samuel N. Emerson, Ye Zheng, Kalee Carson, Patrick J. Cimino, C. Dirk Keene, Jay F. Sarthy, Raphael Gottardo, Kami Ahmad, Steven Henikoff, and Anoop P. Patel. Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. *Nature Biotechnology*, pages 1–6, apr 2021. ISSN 1087-0156. doi: 10.1038/s41587-021-00865-z. URL `https://doi.org/10.1038/s41587-021-00865-z`.

[9] Marek Bartosovic, Mukund Kabbe, and Gonçalo Castelo-Branco. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nature Biotechnology*, pages 1–11, apr 2021. ISSN 1087-0156. doi: 10.1038/s41587-021-00869-9. URL `http://www.nature.com/articles/s41587-021-00869-9`.

[10] Derek H. Janssens, Michael P. Meers, Steven J. Wu, Ekaterina Babaeva, Soheil Meshinchi, Jay F. Sarthy, Kami Ahmad, and Steven Henikoff. Automated CUT&Tag profiling of chromatin heterogeneity in mixed-lineage leukemia. *Nature Genetics*, 53(11):1586–1596, nov 2021. ISSN 15461718. doi: 10.1038/s41588-021-00941-9. URL `https://doi.org/10.1038/s41588-021-00941-9`.

[11] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, aug

2018. ISSN 0028-0836. doi: 10.1038/s41586-018-0414-6. URL http://www.nature.com/articles/s41586-018-0414-6.

[12] Zeinab Navidi, Lin Zhang, and Bo Wang. simATAC: a single-cell ATAC-seq simulation framework. *Genome Biology*, 22(1):1–16, dec 2021. ISSN 1474760X. doi: 10.1186/s13059-021-02270-w. URL https://link.springer.com/articles/10.1186/s13059-021-02270-whttps://link.springer.com/article/10.1186/s13059-021-02270-w.