Modeling fragment counts improves single-cell ATAC-seq analysis

Laura D. Martens[1,2,3], David S. Fischer[2,5], Fabian J. Theis[2,3,4,*] & Julien Gagneur[1,2,3,*]

[1] Department of Informatics, Technical University of Munich, Garching, Germany
[2] Institute of Computational Biology, Helmholtz Center Munich, Neuherberg, Germany
[3] Helmholtz Association - Munich School for Data Science (MUDS), Munich, Germany
[4] Department of Mathematics, Technical University of Munich, Garching, Germany
[5] TUM School of Life Sciences Weihenstephan, Technical University of Munich, 85354 Freising, Germany
* Correspondence to gagneur@in.tum.de or fabian.theis@helmholtz-muenchen.de

Single-cell ATAC-sequencing (scATAC-seq) coverage in regulatory regions is typically binarized as an indicator of open chromatin. However, the implications of scATAC-seq data binarization have never been assessed. Here, we show that the goodness-of-fit of existing models and their applications including clustering, cell type identification, and batch integration, are improved by a quantitative treatment of the fragment counts. These results have immediate implications for scATAC-seq analysis.

Single-cell ATAC-seq (scATAC-seq[1]) is a major method employed to study chromatin regulation in health and diseases. Single-cell ATAC-seq leverages the fact that active cis-regulatory regions are accessible for transcription factors to bind. In the ATAC-seq (assay for transposase-accessible chromatin using sequencing) protocol, the prokaryotic Tn5 transposase is used to insert sequencing adapters into accessible regions of the genome[1,2]. Reads in scATAC-seq data reflect Tn5 insertion events in individual cells[1] (Fig. 1a, example Fig 1b). When analyzing scATAC-seq data, open chromatin regions are generally first called on the pooled data as peaks, i.e., genomic regions with a significant excess of reads compared to background[1,3–5]. Once peaks are called, the reads overlapping each peak are counted for each cell. The resulting count matrix is typically very sparse, with less than 10% non-zero counts[6].

Machine learning modeling of scATAC-seq data is used to support biological investigations of single-cell regulation, including identification of cell types, differentially accessible regions, and inference of transcription factor activity. Independently on the modeling strategy and the application purpose, one guiding criterion for fitting any machine learning model is the choice of

a loss function that captures the goodness-of-fit of a model to the data. Therefore, the data representation and the choice of the loss function are of prime importance for all those models. Due to overall data sparsity, and also because of a conceptualization of chromatin accessibility as a binary state, it is customary to binarize the read count matrix prior to machine learning modeling of scATAC-seq data[4,7–9]. While handling the data quantitatively has been done for some methods[3,5,10,11], there is no systematic investigation of the advantage or disadvantages of binarizing the data.

Here we study the effects of binarization versus count-based models on scATAC-seq data modeling tasks. We based our analysis on two publicly available datasets with cell-type annotation and several batches: (i) a hematopoiesis dataset from Satpathy et al.[12] consisting of flow-sorted and unsorted samples from bone marrow and blood, and (ii) a joint single-cell RNA-seq and ATAC-seq bone marrow dataset from the NeurIPS 2021 Competition Track[13].

We first considered the proportion of peaks above the typical binarization threshold of one read. A substantial number of non-zero peaks had counts of two or even more than two reads (12% and 14.5% for the NeurIPS and Satpathy et al. dataset, respectively, Fig. 1c and Extended Data Fig. 1a). Interestingly, the majority of non-zero peaks had two reads rather than one. When studying the distribution of counts in more detail, we found that it had more peaks with even than uneven counts (Fig. 1c and Extended Data Fig. 1a). This pattern can be explained as an artifact of the count aggregation strategy. Within the 10X Genomics ATAC pipeline[14] and also other pipelines[3,5], the reads or fragment ends instead of the fragments themselves are counted because they mark transposase insertion events. Since scATAC-seq creates paired-end reads, this will mostly result in even counts, whereas uneven counts are only generated if one read pair lies outside the peak region (Fig. 1a,b).

Loss functions are typically based on the choice of a parametric distribution of the data. However, the alternating pattern of odd and even read counts does not suggest a straightforward parameterization. In contrast, fragment counts - estimated here from the existing pipelines by simply aggregating odd and neighboring even reads - showed a regular monotonic decay (Fig. 1d and Extended Data Fig. 1b; Online Methods). When modeling counts, the simplest description is given by the Poisson distribution, for which mean equals variance. We found that the variance of read counts for each region across cells was approximately twice the mean (Fig. 1e and Extended Data Fig. 1c), violating the Poisson assumption and consistent with the fact that reads typically come by two. In contrast, the mean-variance relationship of fragment counts was broadly

consistent with a Poisson distribution (Fig. 1f and Extended Data Fig. 1d) across our two data sets and the entire dynamic range.

Altogether, these results have two implications: First, scATAC-data contains information beyond a binary accessibility state. Second, fragment counts, but not read counts, can be approximately modeled with the Poisson distribution.

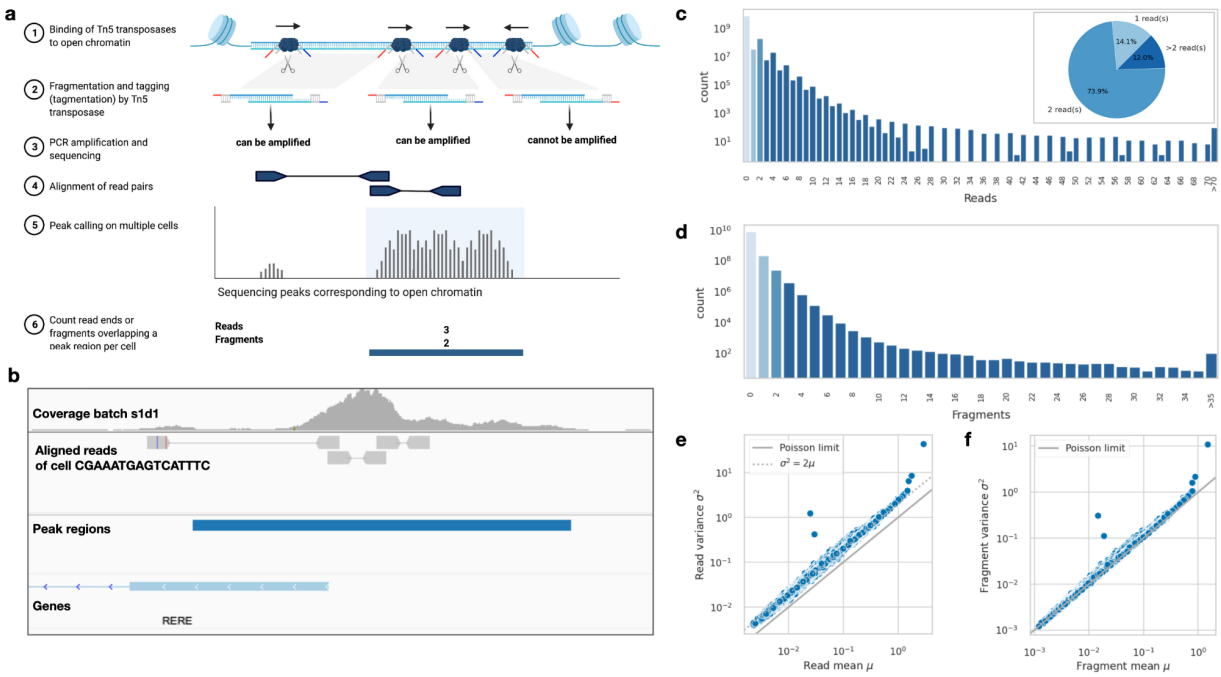

Figure 1. scATAC-seq data is quantitative.

a) Illustrated is the scATAC-seq protocol and count aggregation strategy. Tn5 transposases insert into open chromatin regions, cut the DNA, and attach sequencing adapters (blue and red). Two Tn5 insertions create one fragment with adapters. The orientation of insertion is important as only fragments flanked with two distinct barcodes can be captured and amplified. Fragments are sequenced and aligned to the genome. scATAC-seq peak calling is performed using reads from all cells. Once peak regions are identified, reads or fragments overlapping the peak region are counted for each cell separately. **b)** Genome viewer snapshot of one peak region in the NeurIPS dataset at the promoter of the human gene *RERE* showing multiple insertions in a single cell. The tracks show, from top to bottom, the coverage of one batch used for peak calling, the aligned read pairs of a single cell, the peak region, and genome annotation. The peak region overlaps with five reads and three fragments. **c)** Read count distribution on the entire NeurIPS dataset. The y-axis is on a logarithmic scale. The striking odd/even pattern in read count distribution reflects that reads come in pairs and suggests that fragment counts, rather than reads, should be modeled. Inset: A pie chart showing the percentage of all non-zero peaks with 1, 2, or more than 2 reads. **d)** Distribution of the approximated fragment count does not show an even/odd pattern. **e)** Variance of read counts across cells against mean read counts. Each dot represents one peak region. When fragment ends (reads) are counted, the variance of read counts is about twice the mean (grey dotted line), which is not consistent with a Poisson distribution (solid grey line). The y-axis and the x-axis

are shown on a logarithmic scale. **f)** Variance of fragment counts is approximately equal to the fragment count mean, consistent with a Poisson distribution (solid grey line).

Next, we investigated the effects of modeling the fragment counts rather than a binarized signal on downstream applications. We first investigated latent space learning of sc-ATAC-seq data, a key first step for cell type identification, batch integration and trajectory learning. To this end, we considered as baseline model PeakVI model, a state-of-the-art deep generative model for scATAC-data[8]. PeakVI models binarized ATAC data by learning a Bernoulli probability that a peak in a given cell is accessible. PeakVI corrects for cell-specific effects and region biases by learning cell-specific and region-specific factors. We adapted PeakVI to model Poisson-distributed data in a model we call Poisson VAE (Poisson variational autoencoder). We used the same architecture as PeakVI up to the last layer, where we modeled a Poisson rate parameter for a given peak region in a cell (Online Methods). The total number of fragments per cell varies drastically across cells. We included total fragment count as a pre-computed offset in the loss instead of learning a cell-specific factor as done in PeakVI. Similarly, we tested the effect of using the proportion of non-zeroes in the binary case (Binary VAE) by including it as an offset in the binary loss function (Online Methods).

We first assessed how well the models fitted the two studies. As a common ground, we benchmarked the models for predicting the presence of at least one read. As scores, we used the predicted probability of a region to be open for the binary models and used the predicted probability of non-zero counts for the quantitative models. In this setup, Poisson VAE significantly outperformed PeakVI in reconstructing binarized counts as measured by average precision (NeurIPS: adjusted $P = 1.2 \times 10^{-7}$ and Satpathy et al.: adjusted $P = 6.9 \times 10^{-8}$) (Fig. 2a). Most of the performance boost could be attributed to using the observed total fragment counts rather than predicting it since the binary model (Binary VAE) also showed significantly better reconstruction than PeakVI. In addition, the quantitative model Poisson VAE significantly outperformed the binary model Binary VAE (NeurIPS: adjusted P = 1.3e-06 and Satpathy et al.: adjusted P = 7.1e-03).

We next investigated the implications of these improved fits for downstream applications, starting with batch integration. Since a good integration should reduce technical artifacts between batches while retaining biological variation, we evaluated several integration metrics as previously described[15]. Poisson VAE consistently improved batch integration while better retaining biological signals (Fig. 2b-d and Extended Data Fig. 2 and 3). For example, Poisson VAE better recovers the isolated label ID2-hi myeloid progenitors (only present in 2 of the 13 batches) as measured

by the improved isolated label F1 and isolated label silhouette. Lastly, we assessed the effect of binarization when predicting single-cell ATAC-seq data from matched scRNA-seq data. To this end, we compared to the winning algorithm of the NeurIPS Competition 2021 “Multimodal Single-Cell Data Integration”. We implemented two encoder-decoder models differing only by the loss, binary or Poisson, and in both using the total fragment counts as a given offset (Online Methods). The results are consistent with the above described ATAC-to-ATAC modeling. First, the Poisson-loss model performs significantly better than the Binary-loss model even though it is evaluated for binarized predictions (reconstruction of binarized scATAC-seq data on 10,000 random peaks as measured by root mean squared error, Fig. 2e). Second, a substantial performance boost is reached by using the observed ATAC-seq total fragment counts rather than predicting them from RNA-seq, as both our models largely outperformed the NeurIPS competition winner model (LS_labs), which however only had access to total counts of RNA-seq (Fig. 2e).

In conclusion, we found that scATAC-seq data can be treated quantitatively and that useful information is lost through binarization of the counts. Modeling DNA accessibility in single nuclei quantitatively, rather than as a binary state, is consistent with the fact that to access DNA, transcription factors, just like transposases, have to diffuse through the nucleus, likely reaching distinct chromosome territories and compartments with various efficiency (Fig. 2f). Also, a single genomic position in diploid cells is not necessarily open or close on both alleles synchronously. Moreover, DNA accessibility is not a static but a highly dynamic state. In fact, nucleosome turnover rates[16] are in the order of magnitude of the incubation duration of the transposition reaction[1]. In parallel to our work and in agreement with these findings, a recent preprint showcased that gene expression increases with the number of fragments in accessible regions in single nuclei[17]. We showed that fragment counts, but not read counts, can be modeled with the Poisson distribution. We found the scATAC-seq total fragment counts to be of great importance for both the reconstruction from scATAC data and the prediction of scATAC data from scRNA-seq data. Therefore, we recommend incorporating total fragment counts in models of scATAC-seq data - just as it is typically done for scRNA-seq data - as well as updating corresponding pipelines and benchmarks/competitions. These findings have immediate practical implications for modeling scATAC-seq data since using a Poisson loss over a binary loss comes at no noticeable increase in algorithmic complexity and computing time. Applications of this model to other settings such as multiome integration could be the subject of future work.

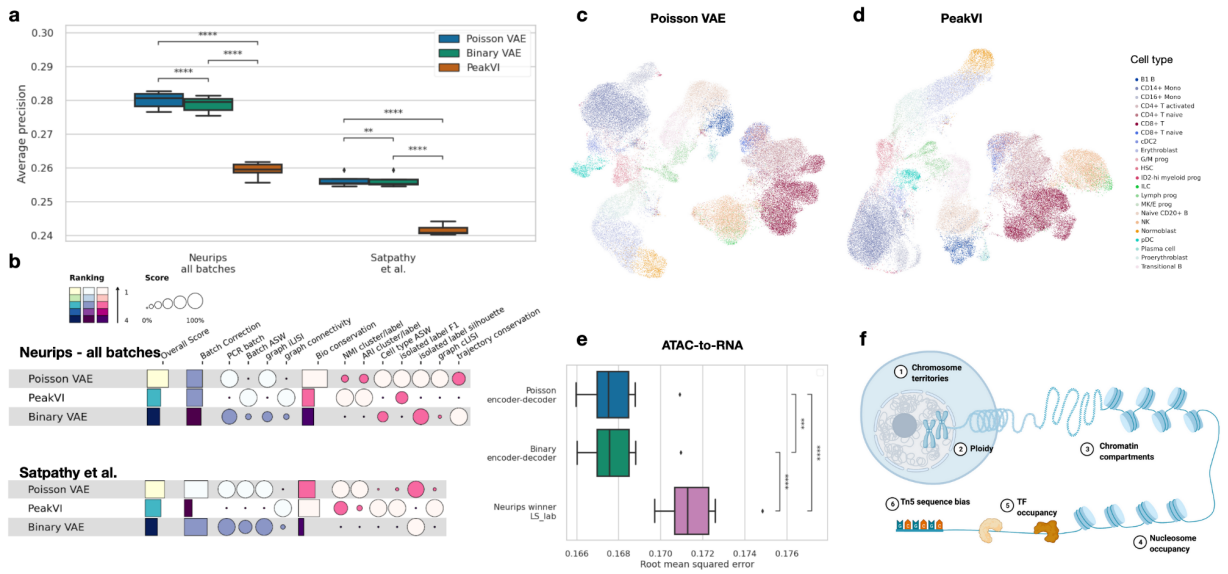

Figure 2. Modeling fragment counts instead of binarized scATAC data improves reconstruction, prediction, and latent representation of cellular profiles. a) Comparison of the Poisson VAE, Binary VAE, and PeakVI model on reconstructing the binarized cell-peak matrix of the NeurIPS and the Satpathy et al. dataset for five cross-validation runs. Poisson VAE and Binary VAE use the observed total fragment count. Poisson rates were transformed to Bernoulli probabilities of generating more than one fragment, and performance was evaluated using average precision. **b)** Comparison of integration accuracy for Binary and Poisson VAE and PeakVI on the NeurIPS dataset and hematopoiesis dataset from Satpathy et al[12]. Shown is the mean of every metric over five cross-validation runs. Metrics are divided into categories of batch correction (blue) and bio-conservation (pink). Individual scores are minimum-maximum scaled between 0 and 1. Overall scores were computed using a 40:60-weighted mean of batch correction and bio-conservation scores, respectively. c, **d)** UMAP of the integrated latent space of all NeurIPS batches, colored by cell type for the Poisson VAE (c) and PeakVI (d) model. **e)** Performance on the NeurIPS challenge subtask of predicting ATAC-seq from gene expression. Each model was trained and evaluated on eleven sets of 10,000 random peaks and evaluated using the same metric as in the original competition RMSE. The best-performing model from the NeurIPS challenge is compared to a binary encoder-decoder model and a Poisson encoder-decoder model (using the observed total fragment count). **f)** Multiple biological factors contribute to DNA accessibility in single cells to be quantitative rather than binary. They include a diploid genome, density of chromatin packaging, nucleosome spacing, TFs in a peak region preventing the Tn5 from binding, and sequence preferences of Tn5. In all boxplots, the central line denotes the median, boxes represent the interquartile range (IQR), and whiskers show the distribution except for outliers. Outliers are all points outside 1.5 times the IQR. *P* values were computed using the paired t-test. *adjusted $P < 0.05$, **adjusted $P < 0.01$, ***adjusted $P < 0.001$, ****adjusted $P < 0.0001$; Benjamini–Hochberg corrected.

Methods

Poisson VAE model

Let $X^{N \times P}$ be a fragment count matrix over N cells and P peak regions. We model the counts with a variational autoencoder:

$$
\begin{aligned}
z_c &\sim \mathrm{Normal}(f_z^{\mu}(x_c), f_z^{\sigma}(x_c)) \\
\rho_{cp} &= f_w^p(z_c, s_c) \\
w_{cp} &= \mathrm{softmax}(\rho_{cp} + r_p) \\
\lambda_{cp} &= \exp(l_c) \cdot w_{cp} \\
x_{cp} &\sim \mathrm{Poisson}(\lambda_{cp})
\end{aligned}
$$

The encoder f_z^{μ}, f_z^{σ} encode the parameters of a multivariate random variable z_c. f_w is a neural network that maps the latent representation z_c concatenated to the batch annotation s_c back to the dimension of peaks. l_c refers to the log-transformed total fragment counts per cell $l_c = \log(\sum_p x_{cp})$. r_p captures a region-specific bias such as the mean fragment count or peak length and is learned directly. w_{cp} is constrained to encode the mean distribution of l_c reads over all peaks by using a softmax activation in the last layer. This means that $\sum_p w_{cp} = 1$.

Binary VAE model

The binary VAE model models binarized counts:

$$
y_{cp} = \begin{cases} 0 \text{ if } x_{cp} = 0 \\ 1 \text{ if } x_{cp} > 0 \end{cases}
$$

The binarized signal was modeled as follows:

$$
\begin{aligned}
z_c &\sim \mathrm{Normal}(f_z^{\mu}(y_c), f_z^{\sigma}(y_c)) \\
\rho_{cp} &= f_w^p(y_c, s_c) \\
p_{cp} &= \sigma(\rho_{cp} + r_p + \tilde{l}_c) \\
y_{cp} &\sim \mathrm{Ber}(p_{cp})
\end{aligned}
$$

We included the proportion of non-zeroes by modeling:

$$
\tilde{l}_c = \sigma^{-1}(1/p \sum_p y_{cp})
$$

Here, σ^{-1} is the logit function. This way $p_{c,p}$ is equal to the mean accessibility of the cell for $\rho_{c,p} = r_p = 0$.

Encoder and decoder functions

The functions f_z^{μ}, f_z^{σ} and the function f_w are encoder and decoder functions, respectively. To be as comparable as possible to PeakVI as implemented in scvi-tools[8,18] (v.0.15.0), we use the same architecture. Specifically, these functions consist of two repeated blocks of fully connected neural networks with a fixed number of hidden dimensions set to the square root of the number of input dimensions, a dropout layer, a layer-norm layer, and leakyReLU activation.

The last layer in the encoder maps to a defined number of latent dimensions n_{latent}.

Adjusting the model to RNA input

To compare the Poisson loss to binary loss in the NeurIPS challenge, we adapted the above-described Poisson VAE and binary VAE models to take scRNA-seq measurements as input. Specifically, they take a matched cells times genes matrix $X_{\text{RNA}}^{N \times G}$ as input with G being the total number of genes. In this case, the latent representation of a cell was derived from the RNA expression vector:

$$z_c \sim \text{Normal}(f_z^{\mu}(x_c^{\text{RNA}}), f_z^{\sigma}(x_c^{\text{RNA}}))$$

Training procedure

We used the default PeakVI training procedure with a learning rate of 0.0001, weight decay of 0.001, and minibatch size of 128 and used early stopping on the validation reconstruction loss. Except for the NeurIPS challenge that provided a test set, we changed the composition of random training, validation, and test set to 80%, 10% and 10% respectively. We computed all evaluation metrics on the left-out test cells.

Input data and preprocessing

Dataset Processing

NeurIPS Dataset

The multiome dataset from the NeurIPS 2021 challenge[13] was downloaded from the AWS bucket *s3://openproblems-bio/public/*. The complete processed dataset is also available on GEO (Accession GSE194122). We did not perform any additional filtering of the data. BAM files for visualization purposes were downloaded from GEO.

Hematopoiesis dataset

The hematopoiesis dataset[12] was downloaded from GEO (Accession GSE129785). Specifically, the processed count matrix and metadata files: scATAC-Hematopoiesis-All.cell-barcodes.txt.gz, scATAC-Hematopoiesis-All.mtx.gz, scATAC-Hematopoiesis-All.peaks.txt.gz. We then filtered the genomic region to only those that are detected in at least 0.1% of the cells in the sample, reducing the data from 571,400 to 134,104 peaks.

Fragment computation

The standard 10X protocol for generating the cell-peaks matrix is to count the fragment ends. As a simple way to estimate fragment counts, we rounded all uneven counts to the next highest

even number and halved the resulting read counts.

Evaluation

Reconstruction metrics

The reconstruction metrics were calculated on the binarized matrix. Poisson rate parameters λ_{cp} were transformed to a Bernoulli probability p_{cp} by computing the probability of getting one or more fragments in a peak for a given cell:

$$p_{cp} = \mathbb{P}(x_{cp} > 0|\lambda_{cp}) = 1 - \mathbb{P}(x_{cp} = 0|\lambda_{cp}) = 1 - e^{-\lambda_{cp}}$$

Average precision

As our reconstruction task is highly imbalanced (only a small fraction of all peaks are accessible), we used the average precision score as implemented in scikit-learn (v.1.0.2) to evaluate the reconstruction. Average precision estimates the area under the precision-recall curve. The score is between 0 and 1 with 1 being the best.

RMSE

As in the NeurIPS challenge, we used the root mean squared error (RMSE) on binarized count matrices to benchmark the RNA-to-ATAC prediction task.

Integration metrics

We used the scib[15] (v.1.0.0) implementation for computing the integration metrics on the latent embedding of the cells. We used all available metrics using default parameters but excluded metrics that were specifically developed for scRNA-seq datasets (highly variable genes score, cell cycle score) and kBET due to its long run time. The trajectory score was only run for the NeurIPS dataset, which had a precomputed ATAC trajectory.

Statistics

scATAC-seq reconstruction

The robustness of the reconstruction metrics was evaluated using five-fold cross-validation. We then employed a two-sided paired t-test to test for significant differences between models and corrected for multiple testing using Benjamini-Hochberg.

NeurIPS competition evaluation

The models in the original NeurIPS competition were evaluated on a single set of 10,000 randomly chosen peaks. To be able to compute statistics on our improvement, we extended this set by ten additional sets of randomly chosen peaks, excluding the original peak set. Statistical

significance was evaluated using a two-sided paired t-test, corrected for multiple testing with Benjamini-Hochberg.

Benchmarking Methods

Hyperparameter optimization

All models were run using the default PeakVI parameters. For the ATAC-ATAC task, we optimized the number of latent dimensions n_{latent} on the validation set for each dataset and model on reconstructing the binary accessibility matrix as measured by average precision. The used range was from 10 to 100 in increments of 10. In the RNA-ATAC prediction task, we optimized the latent dimensions on 10% of the training cells. Since we only trained on 10,000 peaks at a time, the used range was smaller, from 10 to 50 in increments of 10.

NeurIPS competition winner - LS_labs

We reran the winning model LS_labs of the Single NeurIPS Competition 2021 “Multimodal Single-Cell Data Integration” in the category “Predict modality GEX to ATAC” using the published script that we downloaded from GitHub https://github.com/openproblems-bio/neurips2021_multimodal_topmethods. In particular, we used the *script.py* file.

Data availability

Raw published data for the NeurIPS[13] and Satpathy et al.[12] examples are available from the Gene Expression Omnibus under accession codes GSE194122 and GSE129785, respectively.

Code availability

All models, code, and notebooks to reproduce our analysis and figures are available on https://github.com/theislab/scatac_poisson including documentation and examples.

Author contributions

L.D.M conducted the analysis and implemented the models. J.G. and F.J.T. conceived and supervised the project with the help of D.S.F. All authors wrote and contributed to the manuscript. The authors read and approved the final manuscript.

Competing interests

F.J.T. consults for Immunai Inc., Singularity Bio B.V., CytoReason Ltd and Omniscope Ltd, and

has ownership interest in Dermagnostix GmbH and Cellarity. The remaining authors declare no competing interests.

Corresponding authors

Correspondence to Julien Gagneur or Fabian J. Theis.

Acknowledgments

We would like to thank Ignacio Ibarra, Alexander Karollus, and Pedro Tomaz da Silva for valuable feedback and reading of the manuscript. L.D.M acknowledges support by the Helmholtz Association under the joint research school Munich School for Data Science - MUDS" and J.G by the Deutsche Forschungsgemeinschaft (SFB/Transregio TRR267). Fig. 1a is adapted from "ATAC Sequencing", by BioRender.com (2022). Fig. 2f is adapted from "Regulation of Transcription in Eukaryotic Cells", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates.

References

1. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
2. Buenrostro, J., Wu, B., Chang, H. & Greenleaf, W. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al* **109**, 21.29.1-21.29.9 (2015).
3. Stuart, T., Srivastava, A., Lareau, C. & Satija, R. Multimodal single-cell chromatin analysis with Signac. 2020.11.09.373613 (2020) doi:10.1101/2020.11.09.373613.
4. Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1337 (2021).
5. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
6. Li, Z. *et al.* Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. *Nat. Commun.* **12**, 6386 (2021).
7. Bravo González-Blas, C. *et al.* cisTopic: cis-regulatory topic modeling on single-cell ATAC-

335 seq data. *Nat. Methods* **16**, 397–400 (2019).
336 8. Ashuach, T., Reidenbach, D. A., Gayoso, A. & Yosef, N. PeakVI: A deep generative model
337 for single-cell chromatin accessibility analysis. *Cell Rep. Methods* 100182 (2022)
338 doi:10.1016/j.crmeth.2022.100182.
339 9. Xiong, L. *et al.* SCALE method for single-cell ATAC-seq analysis via latent feature extraction.
340 *Nat. Commun.* **10**, 4576 (2019).
341 10. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring
342 transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods*
343 **14**, 975–978 (2017).
344 11. Ji, Z., Zhou, W., Hou, W. & Ji, H. Single-cell ATAC-seq signal extraction and
345 enhancement with SCATE. *Genome Biol.* **21**, 161 (2020).
346 12. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human
347 immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936
348 (2019).
349 13. Luecken, M. D. *et al.* A sandbox for prediction and integration of DNA, RNA, and
350 proteins in single cells. in (2021).
351 14. Cell Ranger ATAC Algorithms Overview. https://support.10xgenomics.com/single-cell-
352 atac/software/pipelines/latest/algorithms/overview.
353 15. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics.
354 *Nat. Methods* **19**, 41–50 (2022).
355 16. Deal, R. B., Henikoff, J. G. & Henikoff, S. Genome-Wide Kinetics of Nucleosome
356 Turnover Determined by Metabolic Labeling of Histones. *Science* **328**, 1161–1164 (2010).
357 17. Miao, Z. & Kim, J. Is single nucleus ATAC-seq accessibility a qualitative or quantitative
358 measurement? 2022.04.20.488960 (2022) doi:10.1101/2022.04.20.488960.
359 18. Gayoso, A. *et al.* A Python library for probabilistic analysis of single-cell omics data. *Nat.*
360 *Biotechnol.* **40**, 163–166 (2022).