A deep generative model of single-cell methylomic data

Ethan Weinberger
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98195
ewein@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98195
suinlee@cs.washington.edu

Abstract

Single-cell DNA methyolme profiling platforms based on bisulfite sequencing techniques promise to enable the exploration of epigenomic heterogeneity at an unprecedented resolution. However, substantial noise resulting from technical limitations of these platforms can impede downstream analyses of the data. Here we present methylVI, a deep generative model that learns probabilistic representations of single-cell methylation data which explicitly account for the unique characteristics of bisulfite-sequencing-derived methylomic data. After initially validating the quality of our model's fit, we proceed to demonstrate how methylVI can facilitate common downstream analysis tasks, including integrating data collected using different sequencing platforms and producing denoised methylome profiles. Our implementation of methylVI is publicly available at https://github.com/suinleelab/methylVI.

1 Introduction

Chromosomal DNA methylation (DNAm) at cytosine residues is known to play a critical role in a broad range of biological processes, including cellular differentiation, genomic imprinting, and X-chromosome inactivation, among others [1–3]. While the majority of methylation in the mammalian genome occurs at CpG sites (i.e., a cytosine followed by a guanine nucleotide), CpH methylation (i.e., methylation at a cytosine followed by a non-guanine nucleotide) is also known to be prevalent in embryonic stem cells and brain tissue [4, 5]. Moreover, abnormalities in DNAm have been implicated in numerous diseases, including cancer, Alzheimer's disease, and diabetes [6–8]. As a result, significant efforts have been made to develop methods for profiling DNAm levels, with bisulfite-sequencing techniques emerging as the gold-standard for DNAm profiling [9].

To better understand heterogeneity in the epigenomic landscape, recent works have focused on developing methods for profiling DNAm at the single-cell level [10–16]. However, analyzing single-cell methylomic data is complicated by multiple nuisance sources of variation, such as differences in sequencing depth, limited coverage of cytosines, and batch effects, that are unrelated to any underlying biological phenomena. Such issues thus necessitate the development of new computational techniques to fully leverage this rich data. Previous works in computational modeling of other single-cell omics modalities have addressed these issues by developing probabilistic latent variable models that model the distinct data generation processes of each modality [17–24]. However, to our knowledge no corresponding works have focused on methylomic measurements. Thus, to address this gap in the literature, here we introduce methylVI, a probabilistic latent variable model of single-cell methylomic data that can facilitate the analysis of single-cell methylomic data.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

2 The methylVI probabilistic model

For a given cell i, single-cell bisulfite sequencing experiments output a set of binary values representing the methylation status at an individual cytosine residue. Due to technological issues, these measurements exhibit highly sparse coverage (i.e., for most cytosines we have missing values), and so in practice it is often preferred to aggregate measurements across larger pre-specified genomic regions (e.g. 100 kilobase pair windows, gene body regions etc.). Moreover, due to their distinct roles [5, 25], CpG and CpH methylation are often analyzed separately. Thus, we may regard the output of a bisulfite sequencing experiment for a given cell i as two pairs of d-dimensional vectors (y_i^G, n_i^G) and (y_i^H, n_i^H). Here y_{ij}^G represents the number of methylated cytosines at CpG sites in region j and n_{ij}^G denotes the total number of profiled CpG sites in region j, and y_{ij}^H and n_{ij}^H are defined analogously for CpH sites. We also let s_i denote the (one-hot-encoded) batch that cell i was collected in.

Let z_i be a k-dimensional set of latent variables with $k << d$ capturing the underlying methylation state of cell i. Previous studies have found that DNAm exhibits significant local spatial correlation. Thus, given z_i and the aggregated measurements described previously, we could potentially model y_{ij} as being drawn from a Binomial distribution, where the probability of methylation is assumed to be constant for all cytosines in a given region. However, previous works have found that methylation read counts generated by bisulfite sequencing technologies exhibit greater dispersion than would be expected based on a Binomial model [26–28]. To account for this overdispersion, we thus choose to model y_{ij} conditioned on the latent variables as being drawn from a Beta-Binomial distribution. Let $f_{\theta^G} : \mathbb{R}^k \to \mathbb{R}^d$ and $f_{\theta^H} : \mathbb{R}^k \to \mathbb{R}^d$ denote neural networks parameterized by θ^G and θ^H that map a cell's latent representations to the full dimensionality of all genomic regions. For CpG features we have the following generative process:

$$
\begin{aligned}
z_i &\sim \mathcal{N}(0, I_d) \\
\mu_{ij}^G &= f_{\theta^G}(z_i, s_i)_j \\
y_{ij}^G &\sim \text{Beta-Binomial}(n_{ij}^G, \mu_{ij}^G, \gamma_j^G),
\end{aligned}
$$

which we define similarly for CpH features. Here we use the mean-dispersion parameterization of the Beta-Binomial distribution, and γ_j^G represents a region-specific dispersion parameter. For notational convenience we let $\nu = \{\theta^G, \gamma^G, \theta^H, \gamma^H\}$ denote the full set of parameters for our generative model.

We now briefly describe the inference procedure for our model. We cannot directly compute the posterior distribution of our latent variables $p(z_i|y_i^G, n_i^G, y_i^H, n_i^H, s_i)$, as computation of the marginal distribution $p(x)$ is intractable. Thus, we learn an approximate posterior $q_\phi(z_i|y_i^G, n_i^G, y_i^H, n_i^H, s_i)$ via variational inference [29], where ϕ denotes the parameters of a neural network as in the variational autoencoder framework [30]. We then learn our model by optimizing the evidence lower bound (ELBO) of $\log p_\nu(y_{1:N})$ with respect to variational parameters ϕ and generative model parameters ν.

3 Evaluation

3.1 Posterior Predictive Checks

We first evaluated methylVI by performing a set of posterior predictive checks (PPCs) similar to those of Gayoso et al. [23]. For our PPCs we considered four single-cell brain cell methylome datasets collected using different bisulfite sequencing platforms: snmC-seq [12], snmC-seq2 [31], sn-m3C-seq [31], and snmCAT-seq [32]. We compared methylVI to factor analysis (FA) and principal component analysis (PCA), which have been applied in previous analyses of single-cell methylome data. These methods take normalized methylation levels for each genomic region as input, and thus we benchmarked them under two normalization schemes: (1) a basic form of normalization in which the average methylation at a genomic region for a given cell is estimated by dividing the number of methylated reads by the coverage at that region for that cell, and (2) the posterior mean estimation procedure implemented in the popular ALLCools [31] single-cell methylome analysis package, in which the parameters of a Beta-Binomial distribution are estimated via the method of moments.

We began by comparing the coefficient of variation (CV) of each genomic region for each model's generated data versus the true data. In particular, for each model we sampled the posterior predictive

	Method	snmC-seq [12]		snmC-seq2 [31]		sn-m3C-seq [31]		snmCAT-seq [32]	
		CpG	CpH	CpG	CpH	CpG	CpH	CpG	CpH
CV	methylVI	**1.23**	**0.71**	**1.33**	**0.64**	**1.96**	**1.48**	**1.27**	**0.54**
	FA (Mean)	3.53	1.92	3.93	2.29	4.25	4.74	3.76	1.53
	FA (ALLCools)	4.42	4.07	5.15	5.05	5.07	6.97	4.84	3.97
	PCA (Mean)	3.33	1.95	3.69	2.27	3.96	4.38	3.55	1.56
	PCA (ALLCools)	4.66	4.13	5.28	5.13	5.19	7.18	5.14	4.13
MWU	methylVI	**15.55**	**15.55**	**18.79**	**18.80**	**16.49**	**16.49**	**16.07**	**16.07**
	FA (Mean)	15.59	15.59	18.82	18.83	16.52	16.57	16.09	16.09
	FA (ALLCools)	15.92	15.95	19.13	19.16	16.85	16.90	16.44	16.47
	PCA (Mean)	15.59	15.59	18.83	18.83	16.53	16.56	16.09	16.09
	PCA (ALLCools)	15.92	15.95	19.14	19.18	16.85	16.89	16.45	16.47

Table 1: Posterior posterior predictive checks. We report the median absolute error of the coefficient of variation (CV) (top) and median log Mann-Whitney U statistic (bottom) separately for CpG and CpH features for each dataset. For both metrics **lower** values are better.

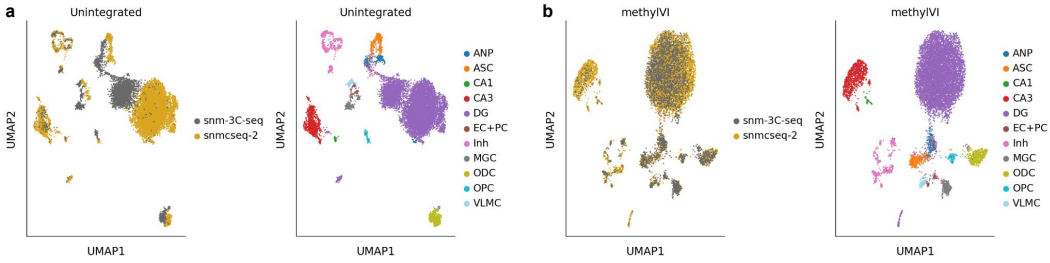

Figure 1: **a-b,** UMAP [33] plots of dentate gyrus methylome data pre-integration (**a**) and after integration with methylVI (**b**) colored by sequencing technology and cell type annotations from [31].

distribution 25 times and calculated the CV for each sample and feature. After averaging over samples we separated the features by methylation type (i.e., CpG versus CpH), and we report the median absolute error between the true and predicted CVs for each methylation type in Table 1. We found that methylVI consistently outperformed baseline models. We also compared the original and replicated data using the Mann-Whitney U (MWU) test statistic. We found (Table 1) that samples produced by methylVI were closer to the true data (i.e., lower MWU) than those generated by baseline models.

3.2 Data Integration

As a further application of the methylVI model, we next evaluated its ability to integrate data from different batches (e.g. datasets collected at different time points, using different technologies, etc.) into a unified latent space via conditioning on provided batch labels s_i for each cell. For this task we considered two sets of methylome measurements from the dentate gyrus gathered as part of a larger mouse brain methylome atlas [31]. These measurements were taken at different time points using different bisulfite sequencing platforms (snmC-seq2 vs. sn-m3C-seq), and exhibit clear separation by batch (Fig. 1a). On the other hand, in methylVI's latent space (Fig. 1b) we find that samples mix across batches and instead separate primarily by cell type.

We benchmarked methylVI's performance on this task with that of three methods originally designed for scRNA-seq data (MNN [34], Scanorama [35], and Harmony [36]) along with ComBat [37], a data integration method designed for bulk assays that has previously been applied to bulk methylome data [38]. To quantify each method's performance, we used the previously established single-cell integration benchmark (scIB) suite of metrics [39]. In short, scIB assesses the quality of integrated representations in terms of both mixing across batches ("Batch correction") and conservation of biological variation ("Bio-conservation"); a successful integration data should achieve good mixing while also conserving biological variation. We found (Table 2) that methylVI substantially outperformed baselines in terms of batch correction, while also comparing favorably on bio-conservation.

Method	Batch Correction	Bio-conservation
methylVI	**0.901**	**0.759**
MNN	0.622	0.341
Scanorama	0.376	0.466
Harmony	0.287	0.242
ComBat	0.342	0.440
Unintegrated	0.237	0.684

Table 2: Data integration performance as measured by the scIB [39] suite of metrics (**higher** is better).

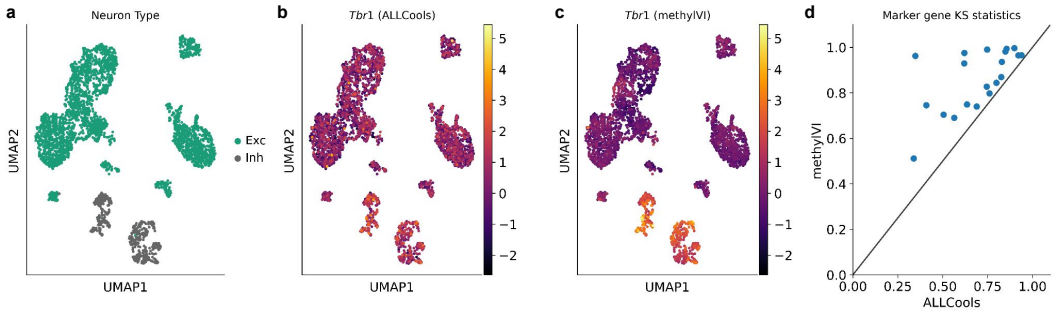

Figure 2: **a,** methylVI's latent representations of excitatory (Exc) and inhibitory (Inh) neurons' methylomes from [12]. **b-c,** Log-transformed normalized *Tbr1* gene body CpH methylation levels as computed by ALLCools (**b**) and methylVI (**c**). **d,** Kolmogorov-Smirnov (KS) test statistics assessing difference in distribution of normalized gene body CpH methylation levels of canonical neuron marker genes between cell types corresponding to the marker gene versus other cell types. CpH methylation values were normalized via the ALLCools pipeline (x-axis) or methylVI (y-axis).

3.3 Data Denoising

As another application of the methylVI model, we used it to construct denoised methylation profiles of single cells. Such a denoised profiles can be used as input to further downstream analysis tasks, such as identifying markers of different cell types or identifying regulatory networks. To construct such a denoised profile, we compute $\mathbb{E}_{q_\phi(z_i|y_i^G,n_i^G,y_i^H,n_i^H,s_i)}[\mu_{ij}^G]$ for CpG methylation for all cells i and regions j and similarly for CpH methylation (i.e., we obtain the latent embedding for each cell i and subsequently compute the mean parameter of the Beta-Binomial distribution for region j in cell i). To validate this denoising procedure, we considered again the mice brain methylome dataset from Luo et al. [12]. Gene expression and gene body CpH methylation levels are known to exhibit an inverse relationship in neurons [5]. In particular, we would expect lower gene body CpH methylation levels for marker genes in cells of the cell type corresponding to that cell type. However, this relationship can be obscured due to noise in the sequencing process. For example, *Tbr1* is a pan-excitatory neuron marker, yet qualitatively it is difficult to distinguish normalized *Tbr1* gene body methylation levels computed using the ALLCools pipeline in excitatory versus inhibitory neurons (Fig. 2a-b). On the other hand, after denoising with methylVI, we observe clear differences (Fig. 2c).

To further illustrate this phenomenon, we considered a set of gold-standard neuron subtype marker genes derived from the literature [40–42]. For each gene, we then applied the Kolmogorov-Smirnov (KS) test to assess the difference of the distribution of normalized gene body CpH methylation levels between cells from a given marker gene's corresponding cell type(s) versus other cell types. In particular, we compared KS statistics computed using ALLCools normalized values versus those of methylVI. As we would expect significantly different distributions of CpH methylation levels between a marker gene's corresponding cell types versus other cell types, higher KS values suggest better recovery of true biological phenomena. We found (Fig. 2d) that methylVI's normalized CpH methylation levels indeed resulted in consistently higher KS values compared to ALLCools.

4 Discussion

Here we introduced methylVI, a deep generative model for the analysis of single-cell methylomic data generated from bisulfite sequencing experiments. Using posterior predictive checks (Section 3.1), we found that methylVI was able to better fit single-cell methylome data compared to baseline linear models employed in previous single-cell methylome analyses. We then applied methylVI to integrate data from different experimental batches (Section 3.2) and to produce denoised methylome profiles. Future work will focus on extending the methylVI model to perform additional downstream analysis tasks (e.g. differentially methylated region testing), and incorporating the methylVI likelihood function into previously developed frameworks (e.g. multiVI [43]) for analyzing multimodal assays.

References

[1] Alexander Meissner, Tarjei S Mikkelsen, Hongcang Gu, Marius Wernig, Jacob Hanna, Andrey Sivachenko, Xiaolan Zhang, Bradley E Bernstein, Chad Nusbaum, David B Jaffe, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770, 2008.

[2] Wolf Reik and Jörn Walter. Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics*, 2(1):21–32, 2001.

[3] Maxim VC Greenberg and Deborah Bourc'his. The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology*, 20(10):590–607, 2019.

[4] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, 2009.

[5] Ryan Lister, Eran A Mukamel, Joseph R Nery, Mark Urich, Clare A Puddifoot, Nicholas D Johnson, Jacinta Lucero, Yun Huang, Andrew J Dwork, Matthew D Schultz, et al. Global epigenomic reconfiguration during mammalian brain development. *Science*, 341(6146):1237905, 2013.

[6] Stephen B Baylin. DNA methylation and gene silencing in cancer. *Nature Clinical Practice Oncology*, 2(Suppl 1):S4–S11, 2005.

[7] Philip L De Jager, Gyan Srivastava, Katie Lunnon, Jeremy Burgess, Leonard C Schalkwyk, Lei Yu, Matthew L Eaton, Brendan T Keenan, Jason Ernst, Cristin McCabe, et al. Alzheimer's disease: early alterations in brain DNA methylation at ank1, bin1, rhbdf2 and other loci. *Nature Neuroscience*, 17(9):1156–1163, 2014.

[8] Amita Bansal and Sara E Pinney. DNA methylation and its role in the pathogenesis of diabetes. *Pediatric Diabetes*, 18(3):167–177, 2017.

[9] Nelly Olova, Felix Krueger, Simon Andrews, David Oxley, Rebecca V Berrens, Miguel R Branco, and Wolf Reik. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biology*, 19(1): 1–19, 2018.

[10] Sébastien A Smallwood, Heather J Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, 11 (8):817–820, 2014.

[11] Chongyuan Luo, Angeline Rivkin, Jingtian Zhou, Justin P Sandoval, Laurie Kurihara, Jacinta Lucero, Rosa Castanon, Joseph R Nery, António Pinto-Duarte, Brian Bui, et al. Robust single-cell DNA methylome profiling with snmC-seq2. *Nature Communications*, 9(1):3824, 2018.

[12] Chongyuan Luo, Christopher L Keown, Laurie Kurihara, Jingtian Zhou, Yupeng He, Junhao Li, Rosa Castanon, Jacinta Lucero, Joseph R Nery, Justin P Sandoval, et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*, 357(6351): 600–604, 2017.

[13] Hanqing Liu, Qiurui Zeng, Jingtian Zhou, Anna Bartlett, B Wang, Peter Berube, Wei Tian, Mia Kenworthy, Jordan Altshul, Joseph R Nery, et al. Single-cell DNA methylome and 3d multi-omic atlas of the adult mouse brain. *bioRxiv*, 2022.

[14] Ryan M Mulqueen, Dmitry Pokholok, Steven J Norberg, Kristof A Torkenczy, Andrew J Fields, Duanchen Sun, John R Sinnamon, Jay Shendure, Cole Trapnell, Brian J O'Roak, et al. Highly scalable generation of DNA methylation profiles in single cells. *Nature Biotechnology*, 36(5): 428–431, 2018.

[15] Ruth V Nichols, Brendan L O'Connell, Ryan M Mulqueen, Jerushah Thomas, Ashley R Woodfin, Sonia Acharya, Gail Mandel, Dmitry Pokholok, Frank J Steemers, and Andrew C Adey. High-throughput robust single-cell DNA methylation profiling with scimetv2. *Nature Communications*, 13(1):7627, 2022.

[16] Sonia Acharya, Ruth V Nichols, Lauren E Rylaarsdam, Brendan L O'Connell, Theodore P Braun, and Andrew C Adey. scimet-cap: High-throughput single-cell methylation analysis with a reduced sequencing burden. *bioRxiv*, pages 2023–07, 2023.

[17] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

[18] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature Communications*, 9(1):284, 2018.

[19] Hanna Mendes Levitin, Jinzhou Yuan, Yim Ling Cheng, Francisco JR Ruiz, Erin C Bush, Jeffrey N Bruce, Peter Canoll, Antonio Iavarone, Anna Lasorella, David M Blei, et al. De novo gene signature identification from single-cell rna-seq with hierarchical poisson factorization. *Molecular Systems Biology*, 15(2):e8557, 2019.

[20] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1):390, 2019.

[21] Tal Ashuach, Daniel A Reidenbach, Adam Gayoso, and Nir Yosef. Peakvi: A deep generative model for single-cell chromatin accessibility analysis. *Cell Reports Methods*, 2(3), 2022.

[22] Tim Stuart, Avi Srivastava, Shaista Madad, Caleb A Lareau, and Rahul Satija. Single-cell chromatin state analysis with signac. *Nature Methods*, 18(11):1333–1341, 2021.

[23] Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature Methods*, 18(3):272–282, 2021.

[24] Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, et al. A python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(2): 163–166, 2022.

[25] Junjie U Guo, Yijing Su, Joo Heon Shin, Jaehoon Shin, Hongda Li, Bin Xie, Chun Zhong, Shaohui Hu, Thuc Le, Guoping Fan, et al. Distribution, recognition and regulation of non-cpg methylation in the adult mammalian brain. *Nature Neuroscience*, 17(2):215–222, 2014.

[26] Egor Dolzhenko and Andrew D Smith. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics*, 15(1):1–8, 2014.

[27] Hao Feng, Karen N Conneely, and Hao Wu. A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Research*, 42(8):e69–e69, 2014.

[28] Omer Weissbrod, Elior Rahmani, Regev Schweiger, Saharon Rosset, and Eran Halperin. Association testing of bisulfite-sequencing methylation data via a laplace approximation. *Bioinformatics*, 33(14):i325–i332, 2017.

[29] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[31] Hanqing Liu, Jingtian Zhou, Wei Tian, Chongyuan Luo, Anna Bartlett, Andrew Aldridge, Jacinta Lucero, Julia K Osteen, Joseph R Nery, Huaming Chen, et al. DNA methylation atlas of the mouse brain at single-cell resolution. *Nature*, 598(7879):120–128, 2021.

[32] Chongyuan Luo, Hanqing Liu, Fangming Xie, Ethan J Armand, Kimberly Siletti, Trygve E Bakken, Rongxin Fang, Wayne I Doyle, Tim Stuart, Rebecca D Hodge, et al. Single nucleus multi-omics identifies human cortical cell regulatory genome diversity. *Cell Genomics*, 2(3), 2022.

[33] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[34] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018.

[35] Brian Hie, Bryan Bryson, and Bonnie Berger. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature Biotechnology*, 37(6):685–691, 2019.

[36] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296, 2019.

[37] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[38] Charlotte S Wilhelm-Benartzi, Devin C Koestler, Margaret R Karagas, James M Flanagan, Brock C Christensen, Karl T Kelsey, Carmen J Marsit, E Andres Houseman, and Robert Brown. Review of processing and analysis methods for DNA methylation array data. *British Journal of Cancer*, 109(6):1394–1402, 2013.

[39] Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1): 41–50, 2022.

[40] Bradley J Molyneaux, Paola Arlotta, Joao RL Menezes, and Jeffrey D Macklis. Neuronal subtype specification in the cerebral cortex. *Nature Reviews Neuroscience*, 8(6):427–437, 2007.

[41] Carl P Wonders and Stewart A Anderson. The origin and specification of cortical interneurons. *Nature Reviews Neuroscience*, 7(9):687–696, 2006.

[42] Ed S Lein, Michael J Hawrylycz, Nancy Ao, Mikael Ayres, Amy Bensinger, Amy Bernard, Andrew F Boe, Mark S Boguski, Kevin S Brockway, Emi J Byrnes, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, 2007.

[43] Tal Ashuach, Mariano I Gabitto, Rohan V Koodli, Giuseppe-Antonio Saldi, Michael I Jordan, and Nir Yosef. Multivi: deep generative model for the integration of multimodal data. *Nature Methods*, pages 1–10, 2023.