bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MultiVI: deep generative model for the integration of multi-modal data

Tal Ashuach[1,2,*], Mariano I. Gabitto[2,3,*], Michael I. Jordan[2,3], and Nir Yosef[1,2,4,5,†]

[*]Contributed Equally
[1]Center for Computational Biology, University of California, Berkeley, CA, USA.
[2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA.
[3]Department of Statistics, University of California, Berkeley, Berkeley, CA, USA
[4]Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, USA
[5]Chan Zuckerberg BioHub, San Francisco, CA, USA
[†]Corresponding author. Email: niryosef@berkeley.edu

Abstract

The ability to jointly profile the transcriptional and chromatin landscape of single-cells has emerged as a powerful technique to identify cellular populations and shed light on their regulation of gene expression. Current computational methods analyze jointly profiled (paired) or individual data modalities (unpaired), but do not offer a principled method to analyze both paired and unpaired samples jointly. Here we present MultiVI, a probabilistic framework that leverages deep neural networks to jointly analyze scRNA, scATAC and multiomic (scRNA + scATAC) data. MultiVI creates an informative low-dimensional latent space that accurately reflects both chromatin and transcriptional properties of the cells even when one of the modalities is missing. MultiVI accounts for technical effects in both scRNA and scATAC-seq while correcting for batch effects in both data modalities. We use public datasets to demonstrate that MultiVI is stable, easy to use, and outperforms current approaches for the joint analysis of paired and unpaired data. MultiVI is available as an open source package, implemented in the scvi-tools framework: `https://docs.scvi-tools.org/`.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 Introduction

The advent of technologies for profiling the transcriptional and chromatin accessibility landscape at a single cell resolution has been paramount for cataloging cellular types and states, identifying important genomic regions, and linking genes to their regulatory elements [1, 2]. However most uses of single-cell RNA-seq (scRNA-seq) [3, 4] and single-cell ATAC-seq (scATAC-seq) [2, 5] have been limited such that a given cell can only be profiled by one technology. Recently, multi-modal single-cell protocols [6] have emerged for simultaneously profiling gene expression and chromatin accessibility in the same cell. This concomitant measurement provides a comprehensive source of information for a more refined categorization of cell states and, ultimately, a better understanding of the regulatory mechanisms that underlie the diversity of states.

A promising application of this emerging multi-modal technology is to provide a way to retrospectively analyze existing datasets of a single modality (scRNA- or scATAC-seq only). Given such a dataset (which are much more common than multi-modal datasets), use of multi-modal data can enable an inference of the "missing" modality and thus reach new conclusions about the diversity and regulation of cell states in a wide array of tissues, cell types and experimental or clinical settings. While some computational methods have emerged that can analyze multi-modal data in isolation [7, 8], the joint analysis of multi-modal and single-modality data necessitates the development of novel computational methods. These methods must be capable of leveraging the power of multi-modal data while accounting for the general caveats of single cell genomics data, most prominently - batch effects, limited sensitivity and noise, and taking into consideration the unique statistical properties of each modality (i.e., quantitative signal for scRNA-seq and a largely binary signal for scATAC-seq).

Here, we introduce MultiVI, a deep generative model for probabilistic and integrative analysis of scRNA-seq, scATAC-seq and paired multi-modal data. MultiVI learns low-dimensional representations of single cells that jointly reflects their chromatin and transcriptional landscapes, while modeling technical confounders and accounting for batch effects. This information can be either observed (for cells obtained from a multi-modal assay) or inferred (for cells obtained from a single-modality assay). Similar to previous deep generative models[9, 10], MultiVI also provides a normalized and batch- corrected view of the high-dimensional data (gene expression, chromatin accessibility; either observed or inferred), along with quantification of uncertainty. We utilize several published datasets to demonstrate the ability of MultiVI to integrate and intepret information across different studies, modalities, and experimental technologies. MultiVI is available in scvi-tools as an open source software, along with detailed documentation and a usage tutorial `https://docs.scvi-tools.org/`.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2 Results

2.1 The MultiVI Model

MultiVI leverages our previously presented variational autoencoding (VAE [11]) models for gene expression (scVI [9]) and chromatin accessibility (PeakVI [10]). Given multi-modal data from a single cell (X) from sample (or batch) S, we divide the observations into gene expression (X_R) and chromatin accessibility (X_A). Two deep neural networks termed 'encoders' learn modality-specific, batch-corrected multivariate normal distributions that represent the latent state of the cell based on the observed data, $q(z_R|X_R, S)$ and $q(z_A|X_A, S)$, from the expression and accessibility observations, respectively. To achieve a latent space that reflects both modalities, we penalize the model so that the distance between the two latent representations is minimized and then estimate the integrative cell state $q(z|X_R, X_A, S)$ as the average of both representations. For "unpaired" cells, i.e., cells for which only one modality is available, the cell state is drawn directly from the representation for which data is available (i.e., z_R or z_A).

In the second part of the model, observations are probabilistically generated from the latent representation using two modality-specific deep neural networks termed 'decoders'. Simialr to our previous generative models for gene expression (scVI) and accessibility (PeakVI), the model assumes that the RNA expression data is drawn from a negative binomial distribution, and the accessibility data fits a Bernoulli distribution. The likelihood of the model is computed from both modalities for paired (multi-modal) cells, and only from the respective modality of unpaired cells. Finally, during training, we include an adversarial component which penalizes the model if cells from different modalities are overly separable in latent space.

This two-part architecture enables MultiVI to achieve several goals: first, it leverages the paired data to learn a low-dimensional representation of cell state, which reflects both data types. Second, it allows cells for which only one modality is available to be represented at the same (joint) latent space. Finally, the 'decoding' part of the model provides a way to derive normalized, batch-corrected gene expression and accessibility values for both the multi-modal cells (i.e., normalizing the observed data) and for unpaired cells (i.e., imputing unobserved data; see Figure 1A and methods).

2.2 MultiVI integrates paired and unpaired samples

To study how well MultiVI integrates paired and single-modality data into a common low-dimensional representation, we inspected the outcome of artificially unpairing a jointly profiled dataset. Using a multi-modal PBMC dataset from 10X genomics, a randomly selected set of peripheral blood mononuclear cells (PBMCs; at varying rates, from 1% to 99%), are made unpaired such that each cell appears twice: once with only gene expression data, and once with only chromatin accessibility data. This action resulted in a heterogeneous data set containing three sets of cells: one set has both modalities available, a second

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

set has only RNA-seq information, and the third set of cells has only ATAC-seq information present.

Since, to the best of our knowledge, no published methods offer a principled way to integrate paired multi-modal data with unpaired expression or chromatin data, we attempted to use the Seurat V4 code base[12] to achieve this, and compared the results to those obtained with MultiVI. We applied Seurat using three different approaches: (1) *gene activity*: we converted the ATAC-seq data of the accessibility-only cells to gene activity scores (using the *signac* procedure), and then integrated all the cells using the gene- level data (i.e., gene scores when RNA-seq is not available, or gene expression when RNA-seq is available); (2) *imputed*: we followed the steps in (1) and then used Seurat to impute the RNA expression values for the accessibility-only cells. This is done by averaging over nearby cells in the integrated space for which RNA-seq is available (methods). The data from the accessibility-only cells was then re-integrated with the remaining cells using the imputed RNA expression values instead of the gene scores; (3) *WNN*: using weighted nearest neighbor graphs, which leverages information from both modalities to create a joint representational space, then project single-modality data onto this space (methods).

We ran all methods on the artificially unpaired datasets and compared their latent representations, with the exception of the WNN-based approach and the 99% unpaired dataset, which failed to produce results due to the low number of paired cells (Figure 2A-B, Supplemental Figure 1). We first quantified the mixing abilities of the different approaches, by calculating the local inverse Sampson's index (LISI) score described by [13]. Briefly, for each unpaired cell the fraction of neighbors among the K nearest single-modality neighbors that are of the same modality (expression or accessibility), for varying values of K, normalized by the overall fraction of that modality. This results in an enrichment score, with 1 being perfect mixing (Figure 2C). We found that MultiVI outperforms all other approaches in low rates of unpaired cells (below 50%), but the mixing performance decreases when the rate of unpaired cells increases. Conversely, the Seurat-based imputation approach (unlike the other two Seurat-based approaches) maintains high mixing performance across all levels of unpaired cells. This result is expected, since each accessibility-only cell is represented by an average over cells for which RNA-seq data is available, and that have similar gene expression profiles (i.e., a local neighborhood in a transcriptome- based space). It does not, however, indicate whether these representations are indeed accurate.

Next, we examine the accuracy of the inferred latent space. To measure how well each method captures the true biological state of a cell, we took advantage of the ground truth information contained in our artificially unpaired datasets. For the unpaired cases, we have two distinct representations of the same cell: one based solely on the expression profile and the other solely on the chromatin landscape. Ideally, the two representations would be situated closely in the latent representation, as both capture the same biological state. To measure this, we looked at the distances between the two representations of the unpaired cells in the latent space created by each method. To account for the varying

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

scales of different latent spaces, we used the rank distance (the minimal K for which the two representations are within each other's K nearest neighbors, averaged across all cells; methods, Figure 2D). In this experiment, we found that MultiVI maintains the multi-modal mixing accuracy substantially better than competing methods, and that all methods have a deteriorating performance as the level of unpaired cells increases.

As another test of accuracy, we evaluated the extent to which the stratification of cells into sub-populations remains consistent when most of the data is paired, and when paired data is only partially available. Ideally, if cells are deemed to be at the same cluster based on multi-modal data, they should still be at the same cluster even if they only have information available for one modality. To test this, we clustered the cells in the fully paired and the 75% unpaired data sets (Supplemental Figure 2, methods). Then, we assessed how well the identified clusters in both data sets coincide by measuring the cluster correspondence between data sets (Figure 2E,F). To quantify this consistency, we computed the Adjusted Rand Index (ARI) for both methods, indicating a more faithful representation of the data in MultiVI (MultiVI ARI 0.81, Seurat Imputation ARI 0.66, Figure 2E,F).

Taken together, these results show that MultiVI efficiently integrates unpaired scRNA and scATAC data while capturing the true biological state of each cell, while the Seurat-based approaches either ineffectively mix the modalities, or mix them well but less accurately.

2.3 Integration of Independent Data

Our previous analyses rely on artificially unpaired data, where our model benefits from all data fundamentally being generated in a single batch and by a single technology. While allowing for more accurate benchmarks, this does not reflect real-world situations in which it is desired to integrate data sets that were generated at different batches or even different studies, while possibly using different technologies. We therefore sought to demonstrate MultiVI on a set of real-world data. We collected three distinct datasets of PBMCs: 1) Multi-modal data from the 10X dataset we used previously; 2) ATAC-seq from a subset of Hematopoeisis data generated by Satpathy et al[14], containing multiple batches of PBMCs as well as cell-type specific (FACS-sorted) samples; 3) RNA-seq PBMC data generated by various single cell technologies for a benchmarking study by Ding et al [15]. The datasets were processed to create a set of shared features (genes, loci), and annotations were collected from both Satpathy et al and Ding etl al datasets and combined to a shared set of cell type labels (methods). The resulting dataset has 63382 (53%) ATAC-only cells from Satpathy et al, 44615 (37%) RNA-only cells from Ding et al, and 12012 (10%) jointly profiled cells from 10X.

To gauge the extent of batch effects in this data, we first ran MultiVI without accounting for the study of origin of each sample or to its specific technology (which varies between the RNA-seq samples from Ding et al). With this application, we found substantial batch effects, both between different samples in

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the chromatin accessibility data and between technologies in the gene expression data (Supplemental Figure 3). We then reanalyzed the data, this time configuring MultiVI to correct batch effects and technology-specific effects within each dataset (methods). The resulting, corrected, joint latent space mixes the three datasets well (Figure 3A), while accurately matching labeled populations from both datasets (Figure 3B). MultiVI achieves this while also correcting batch effects within the Satpathy data and technology-specific effects within the Ding data (Figures 3C-D). To better examine the correctness of the integration, we examined the set of labelled cells from the two single-modality datasets (FACS-based labels from Satpathy and manually annotated cells from Ding). For each cell, we examined its 100 nearest neighbors that came from the other modality, and summarized the distribution of labels of those neighbors. We find a clear agreement between the labels of each cell and the labels of it's neighbors, with some mixing among related cell types (Figure 3E). This analysis demonstrates that MultiVI is capable of creating a biologically meaningful representation that mixes data from different sources and modalities, and can be used to leverage previously acquired data in conjunction with new multi-modal data.

2.4 Probabilistic Data Imputation

The probabilistic nature of MultiVI allows the model to impute missing observations, modalities, and generate uncertainty estimates for imputed values. To demonstrate MultiVI's imputation abilities, we resorted to the 10X data set where 75% of the cells were artificially unpaired (as in Figure 2). We used MultiVI to impute the missing modality for the unpaired cells and found that for both modalities, the imputation had high correspondence to the observed values (Figures 4A-B). Specifically, we observe Spearman correlation 0.57 between the imputed expression values and the observed data (taking the raw values, scaled by library size), and an area under the precision-recall curve (PRAUC) of 0.41 for the accessibility data (taking the raw, binary signal). Since the raw data can be largely affected by low sensitivity, we also calculated the correlation between the imputed values and a smoothed version of the data (obtained with a method different of MultiVI; methods), where the signal is average over similar cells (separately for ATAC-seq and RNA-seq), thus mitigating this issue. As expected, we see a higher level of correspondence between the imputed values and this corrected version of the raw data (Spearman correlations 0.8, 0.86 for accessibility, expression respectively; Figure S4A-B).

Next, we focus our analysis on uncertainty estimation for the imputed accessibility values. We measured the uncertainty of the model for each imputed accessibility value by sampling from MultiVI's generative model (methods) and found a strong relationship between the estimated uncertainty and the error of each data point ($(\text{imputed} - \text{observed})^2$), indicating that the model is indeed less certain of predictions that are farther from the unobserved "ground- truth" values (Figure 4C). Equivalent analysis for expression imputations is hindered by the high correlation between the average expression and both the measured error and the uncertainty of the imputed results.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Interestingly, we identified a small subset of values (roughly 0.5% of observations) for which we have high confidence imputations that are associated with high error, when comparing to the unobserved raw accessibility data (Figure 4C, green square). In the case of chromatin accessibility, these high-confidence-high-error imputations correspond to cases where the model confidently predicts the opposite of the actually observed value (Figure 3D). To investigate the source of these errors, we inspected the same cases when comparing the imputed values to the smoothed accessibility estimates (methods). We found that many of these regions were detected as inaccessible in the raw data, but predicted to be accessible by MultiVI, and vice-veras. Interestingly, the smoothed data agrees with the MultiVI predictions - namely, observations that were predicted as accessible tend to be open in highly-similar cells, and observations that were predicted as inaccessible tend to be closed in high-similar cells (Figure 4E). This indicates that these high-confidence-high-error values may correspond to false-negatives and false-positives in the raw data.

As a specific example for imputation, we highlight the T-cell marker gene CD3G. While the observed expression and the observed accessibility of the region containing the transcription start site (TSS) of the gene show high noise and sparsity, the imputed values are highly consistent and clearly mark the T-cell compartment of the latent space (Figure 4F).

Overall, these results show that MultiVI is capable of imputing missing observations, and quantifying the uncertainty for each value, allowing the user to then determine which imputed values are reliable for downstream analyses and which are not. Importantly, since imputing values only requires a latent representation, MultiVI can seamlessly impute missing modalities in a well-integrated space.

3 Discussion

MultiVI is a deep generative model for the integrated analysis of single cell gene expression and chromatin accessibility data. MultiVI uses jointly profiled data to learn a multi-modal model of the data and to relate measurements of individual modalities on the same population of cells. The model accounts for various technical sources of noise and can correct additional sources of unwanted variation (e.g batch effects). MultiVI learns a rich latent representation of the data coalescing information present in each individual data type, which can be used for further single-cell sequencing analysis.

Recent algorithms for the analysis of multi-modal data were developed to process paired datasets, in which both modalities have been profiled at the same cell [8, 7]. These algorithms handle multi-modal data, but lack the ability to integrate single modality datasets into the same analysis. While this task is possible to achieve with the Seurat code base [12], the respective methods we utilized here were not specifically designed to this end, and their performance was not tested for this task. Addressing this need for an integration method, we have shown that MultiVI effectively combines unpaired scRNA and scATAC

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

data with multi-modal single-cell data, generating a robust and meaningful representation of the cells' state that captures information about both their transcriptome and epigenome. This joint representation is achievable even when the amount of paired data is minimal, thus opening exciting opportunities for future analysis. First, MultiVI and similar methods will enable reanalysis of the large compendia of available unpaired datasets (representing the the majority of existing data) with relatively small additional paired data, opening the way for more comprehensive characterization of cell state. Second, it facilitates cost-effective designs for future studies, in which only a subset of samples need to be profiled with the (more costly) multi-modal protocol. Lastly, the probabilistic nature of MultiVI enables data imputation and uncertainty quantification on imputed values, opening the way for downstream statistical analyses that hinge on unobserved values, such as differential accessibility in RNA- only data, or cell state annotation based on marker genes in ATAC- only data.

In summary, MultiVI is able to seamlessly process single and multi-modal data, integrate different chromatin and transcriptional batches, and create a rich joint representation harnessing all available information. The representations created by MultiVI are robust to different composition of multi-modal data sets. Due to properly modeling each individual modality, MultiVI accounts for technical confounders and batch effects. MultiVI is a probabilistic graphical model, given its probabilistic nature, differential expression and accessibility analysis can be performed. Lastly, MultiVI is implemented in the scvi-tools framework [16], making it easy to configure, train, and use.

Declarations

Acknowledgements

We thank Adam Gayoso for assistance on model implementation in scvi-tools. We thank Christina Usher for assistance with visualizations.

Author contributions

TA, MIG and NY conceived of the model and designed the analyses. TA and MIG implemented the model with input from AG. TA performed the analyses. NY and MJ supervised the work. TA, MIG, MJ and NY wrote the manuscript.

Data and Software Availability

All data used in this manuscript is publicly available via the original publications and releases. Intermediate data, trained models used in this manuscript, and the notebooks to generate the figures in this manuscript, are all posted and available on zenodo: `10.5281/zenodo.5225858`.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conflicts of interests

All authors declare that they have no competing interests.

References

[1] Bosiljka Tasic et al. "Adult mouse cortical cell taxonomy revealed by single cell transcriptomics". en. In: *Nat. Neurosci.* 19.2 (Jan. 2016), pp. 335–346. ISSN: 1097-6256. DOI: 10.1038/nn.4216. URL: https://www.nature.com/articles/nn.4216.

[2] Jason D Buenrostro et al. "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide". en. In: *Curr. Protoc. Mol. Biol.* 109 (Jan. 2015), pp. 21.29.1–21.29.9. ISSN: 1934-3639, 1934-3647. DOI: 10.1002/0471142727.mb2129s109. URL: http://dx.doi.org/10.1002/0471142727.mb2129s109.

[3] Fuchou Tang et al. "mRNA-Seq whole-transcriptome analysis of a single cell". en. In: *Nat. Methods* 6.5 (May 2009), pp. 377–382. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1315. URL: http://dx.doi.org/10.1038/nmeth.1315.

[4] Diego Adhemar Jaitin et al. "Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types". en. In: *Science* 343.6172 (Feb. 2014), pp. 776–779. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1247651. URL: https://science.sciencemag.org/content/343/6172/776.

[5] Jason D Buenrostro et al. "Single-cell chromatin accessibility reveals principles of regulatory variation". en. In: *Nature* 523.7561 (June 2015), pp. 486–490. ISSN: 0028-0836. DOI: 10.1038/nature14590. URL: https://www.nature.com/articles/nature14590.

[6] Junyue Cao et al. "Joint profiling of chromatin accessibility and gene expression in thousands of single cells". en. In: *Science* 361.6409 (Sept. 2018), pp. 1380–1385. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aau0730. URL: http://dx.doi.org/10.1126/science.aau0730.

[7] Ricard Argelaguet et al. "MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data". en. In: *Genome Biol.* 21.1 (May 2020), p. 111. ISSN: 1465-6906. DOI: 10.1186/s13059-020-02015-1. URL: http://dx.doi.org/10.1186/s13059-020-02015-1.

[8] Rohit Singh et al. "Schema: metric learning enables interpretable synthesis of heterogeneous single-cell modalities". en. In: *Genome Biol.* 22.1 (May 2021), p. 131. ISSN: 1465-6906. DOI: 10.1186/s13059-021-02313-2. URL: http://dx.doi.org/10.1186/s13059-021-02313-2.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[9] Romain Lopez et al. "Deep generative modeling for single-cell transcriptomics". en. In: *Nat. Methods* 15.12 (Dec. 2018), pp. 1053–1058. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-018-0229-2. URL: http://dx.doi.org/10.1038/s41592-018-0229-2.

[10] Tal Ashuach et al. "PeakVI: A Deep Generative Model for Single Cell Chromatin Accessibility Analysis". en. Apr. 2021. URL: https://www.biorxiv.org/content/10.1101/2021.04.29.442020v1.

[11] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: (Dec. 2013). arXiv: 1312.6114v10 [stat.ML]. URL: http://arxiv.org/abs/1312.6114v10.

[12] Yuhan Hao et al. "Integrated analysis of multimodal single-cell data". en. In: *Cell* 0.0 (May 2021). ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2021.04.048. URL: http://www.cell.com/article/S0092867421005833/abstract.

[13] Ilya Korsunsky et al. "Fast, sensitive and accurate integration of single-cell data with Harmony". en. In: *Nat. Methods* 16.12 (Nov. 2019), pp. 1289–1296. ISSN: 1548-7091. DOI: 10.1038/s41592-019-0619-0. URL: https://www.nature.com/articles/s41592-019-0619-0.

[14] Ansuman T Satpathy et al. "Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion". en. In: *Nat. Biotechnol.* 37.8 (Aug. 2019), pp. 925–936. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-019-0206-z. URL: http://dx.doi.org/10.1038/s41587-019-0206-z.

[15] Jiarui Ding et al. "Systematic comparison of single-cell and single-nucleus RNA-sequencing methods". en. In: *Nat. Biotechnol.* 38.6 (June 2020), pp. 737–746. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-020-0465-8. URL: http://dx.doi.org/10.1038/s41587-020-0465-8.

[16] Adam Gayoso et al. "scvi-tools: a library for deep probabilistic analysis of single-cell omics data". en. Apr. 2021. URL: https://www.biorxiv.org/content/10.1101/2021.04.28.441833v1.

[17] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877. DOI: 10.1080/01621459.2017.1285773. eprint: https://doi.org/10.1080/01621459.2017.1285773. URL: https://doi.org/10.1080/01621459.2017.1285773.

[18] *LiftOver Utility*. URL: https://genome.ucsc.edu/cgi-bin/hgLiftOver.

[19] *Integrating scrna-seq and scatac-seq data*. URL: https://satijalab.org/seurat/articles/atacseq_integration_vignette.html.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4 Methods

4.1 The MultiVI Model

MultiVI inherits generative models describing chromatin accessibility and transcriptional observations from scVI [9] and peakVI [10]. Briefly, Let $X_R \in \mathbb{N}_0^{C \times G}$ be a scRNA-seq genes-by-cell matrix with C cells and G genes, where $x_R^{cg} \in \mathbb{N}_0$ is the number of reads from cell c that map to gene g. Let $X_A \in \mathbb{N}_0^{C \times J}$ be a scATAC-seq region-by-cell matrix with C cells and J regions, where $x_A^{cj} \in \mathbb{N}_0$ is the number of fragments from cell c that map to region j.

MultiVI models the probability of observing x_{cj} counts in a gene by using a negative binomial distribution,

$$x_R^* \sim \text{NegativeBinomial}\left(\ell_c \rho_{cg}, \theta_g\right) \tag{1}$$

where ℓ_c is a scaling factor that captures cell-specific biases (e.g library size), $\rho_c g$ is a normalized gene frequency and θ_g models the per gene dispersion. The probability of observing a region as accessible is modeled with a Bernoulli distribution,

$$x_A^* \sim \text{Ber}\left(\ell_c p_{cj} r_j\right) \tag{2}$$

where p_{cj} captures the true biological heterogeneity; r_j captures region-specific biases (e.g width, sequence). In both observational models, the scaling factor the region-specific and the per gene dispersion parameters are inferred from data using deep neural networks (this is in contrast to the original implementation of SCVI in which library size was modelled using a lognormal distribution).

Next, for each cell, normalized gene frequencies ρ_{cg} and biological heterogeneity p_{cj} are estimated using a latent representation as in VAE[11]. Briefly, each modality is assign their own latent representation, a isotropic multivariate normal distribution $Z_c^A \sim \text{MVN}(0,1)$ and $Z_c^R \sim \text{MVN}(0,1)$. Then, with the purpose of bringing both representations together, they are combined by taking their average $Z_c = \frac{Z_c^A + Z_c^R}{2}$. This merged representation is then used to decode both model parameters, $p_{c.} = f(Z_A)$ and $\rho_{c,} = g(Z_A)$.

4.2 MultiVI Inference Model

We use variational inference [17] to compute posterior estimates of model parameters using the following variational approximation:

$$q(z^R, z^A, r, \ell, \theta | x, y) = q(z^R|x) q(z^A|y) \delta_{\ell *} \delta_{\theta *} \delta_{r *} \tag{3}$$

where delta distribution δ highlight the fact that parameters are inferred from the data as point estimates. The cell-specific factor ℓ_c is computed from the input data for cell c via a deep neural network $f_\ell : \mathbb{N}_0^K \rightarrow [0,1]$. The region-specific factor r_j, since it is optimized across samples, is stored as a K-dimensional tensor, used and optimized directly. In the case of each latent representation, two

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

encoders are computed as $h_z^{Transc} : \mathbb{N}_0^K \rightarrow \left(\mathbb{R}^D, \mathbb{R}^D\right)$ and $h_z^{Chrom} : \mathbb{N}_0^K \rightarrow \left(\mathbb{R}^D, \mathbb{R}^D\right)$ where each of them computes the distributional parameters of a D-dimensional multivariate normal random variable: $Z \sim MVN\left(h_z\left(x_c\right)_1, h_z\left(x_c\right)_2\right)$.

Using the variational approximation, the evidence lower bound (ELBO) is computed and optimized with respect to the variational and model parameters using stochastic gradients. To enforce the similarity between chromatin and transcription latent representations, we add to the ELBO a term that penalizes the distance between representations using a symmetric KL divergence between distributions $d(Z_c^A, Z_c^R) = \text{symmKL}(q(z_c^A), q(z_c^R))$.

5 Benchmarking and Evaluation

5.1 Dataset Preprocessing

The 10x multiomic unsorted PBMC dataset was downloaded from the company website. For artificial unpairing analyses, the processed peak-by-cell matrix was downloaded and filtered to remove features that are detected in fewer than 1% of the cells. For the mixed-source PBMC dataset, the fragment file was downloaded and reprocessed using CellRanger-ARC (v2.0.0) with the Satpathy hg38 peaks. The Satpathy dataset was downloaded from GEO (Accession GSE129785); specifically the processed peak-by-cell 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 data to only include peaks that were detected in at least 0.1% of the data, and lifted those peaks over from the hg19 to the hg38 genome reference using the UCSC liftover utility [18]. The Ding dataset was downloaded from GEO (Accession GSE132044); specifically the pbmc data: $pbmc_hg38_count_matrix.mtx.gz, pbmc_hg38_cell.tsv.gz, p$ Matching cell type annotation was downloaded from SCP (Accession SCP424). After preprocessing, the reanalyzed 10x dataset was combined with both single-modality datasets, and the features were filtered to remove features (either genes or peaks) that were detected in fewer than 1% of the cells.

5.2 RNA-based Seurat integration

This integration modality, disregards multiomic information and only RNA information is considered from multiome cells. Briefly, RNA information is first integrated and then, chromatin accessibility is integrated using gene activity scores (*RNA-based* method) or RNA imputed values (*RNA-based Imputed* method).

In more detail, cells were separated into three different datasets, multiomic cells (using only expression data), rna-only cells and atac-only cells. Seurat objects were created for multiome and rna-only data, and were then normalized, scaled, and the first 50 principal components are calculated. For atac-only cells, a Seurat object was created, gene activity scores were calculated, scaled, and principal components were computed. To integrate the three datasets, integration anchors (using *FindIntegrationAnchors*) were calculated and the data

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

was then integrated (using *IntegrateData*). The *RNA-based* method uses gene activity scores as representative values from the atac-only cells. The *RNA-based Imputed* method includes an additional step in which RNA imputed values are calculated from gene activity scores by running *FindTransferAnchors* and *TransferData*.In this integration method, RNA imputed values are used as representative values from atac-only cells. Finally, integrated data was then scaled and principal components were calculated to generate the final latent space. Across these integration methods, we followed the standard recommended procedure for analyzing data with Seurat given in their tutorials [19].

5.3 WNN-based Seurat Integration

This approach aims to leverage information from both modalities (chromatin accessibility and expression values), using the newly described weighted nearest neighbors approach from Seurat V4 [12]. We first created a weighted nearest neighbor graph using multiomic information and then project chromatin and transcriptional information onto this.

We begin by separating cells in unpaired datasets into three different datasets, multiomic cells (with both expression and chromatin data), rna-only, and atac-only. First, multiome latent representation is found by calculating SC transform and principal components on the expression data and latent semantic analysis (TF-IDF decomposition followed by SVD) on the chromatin data. Next, multimodal neighbors and the first 50 supervised PCA are calculated. To merge rna only and atac only data to multiome representation, transfer anchors (FindTransferAnchors) are computed on rna only data and gene activity scores on atac only and each datasets is integrated using IntegrateEmbeddings function. Finally, datasets and dimensionality reductions are merged and umap is visualized using the merged information.

5.4 Neighbor Rank Distance Calculation

For artificially unpaired cells, each cell has two unpaired representations in the latent space. Given cell c with representations c_a and c_b, let $S(c_a, K)$ be the set of K nearest neighbors to c_a. We then define $\delta(c_a, c_b)$ as the minimal K for which cell c_b is among the K nearest neighbors of cell c_a: $\min\{k : c_b \in S(c_a, k)\}$.

5.5 Enrichment Score Calculation

Enrichment scores were computed as they were in our previous work[10], and similarly to the LISI scores described in the Harmony paper[13]. Briefly, given a latent representation R, an integer k, and the modality labels (expression, or accessibility) L, we compute $G_{R,k}$ the K-nearest neighbor graph from R with k neighbors. Using $G_{R,k}$, we compute for each cell the proportion of neighbors that share the same modality: $s_i = \frac{1}{k}\sum_{j \in G_{R,k}(i)} \mathbb{1}(L_i = L_j)$. The enrichment score is the average score across all cells, $\bar{s}$, normalized by the expected score

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for a random sample from the distribution of labels: $E[s] = \sum_{\ell \in \{L\}} p_\ell^2$, with p_ℓ being the proportion of each modality.

5.6 Estimating Imputation Uncertainty

We estimate the uncertainty of the model for each imputed value by sampling from the latent space (n=15) and computing the standard deviation of the imputed values for each observation. More consistent predictions correspond to less uncertainty.

5.7 KNN-based estimate of accessibility

To estimate accessibility without using MultiVI, we computed a lower-dimensional representation of the data using Latent Semantic Analysis (LSA, top 30 components), then for each cell we computed the average accessibility profile of the 50 nearest neighbors in the LSA space. This creates a smooth estimate of accessibility using highly-similar cells, mitigating the effect of false observations.

5.8 Expression Smoothing

Expression smoothing was achieved by taking the top 30 principle components of the expression data (computed with PCA), computing the K-nearest neighbors graph (for $K = 50$) and averaging the expression values of the neighbors for each cell (scaled by library size).

6 Figures

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

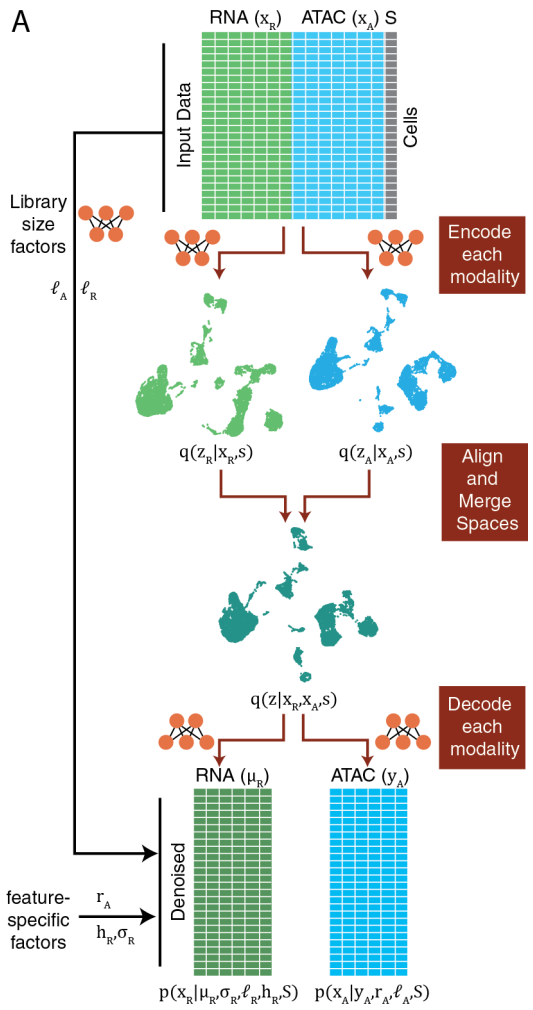

Figure 1: MultiVI Model Overview. Conceptual model illustration in which input data (top) consists of either chromatin accessibility (ATAC), gene expression (RNA) or both data types (Multiome). Variable S represents experimental covariates, such as batch or experimental condition. Each data modality is encoded into modality-independent latent representations (using neural network encoders) and then, these representations are merged into a joint latent space. The joint latent representation is used to estimate (decode) the input data together with chromatin region-specific effects (r_A), gene-specific dispersion (σ_R), cell-specific effects (ℓ_A, ℓ_R), accessibility probability estimates (Y_Z) and mean gene expression values (μ_R).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

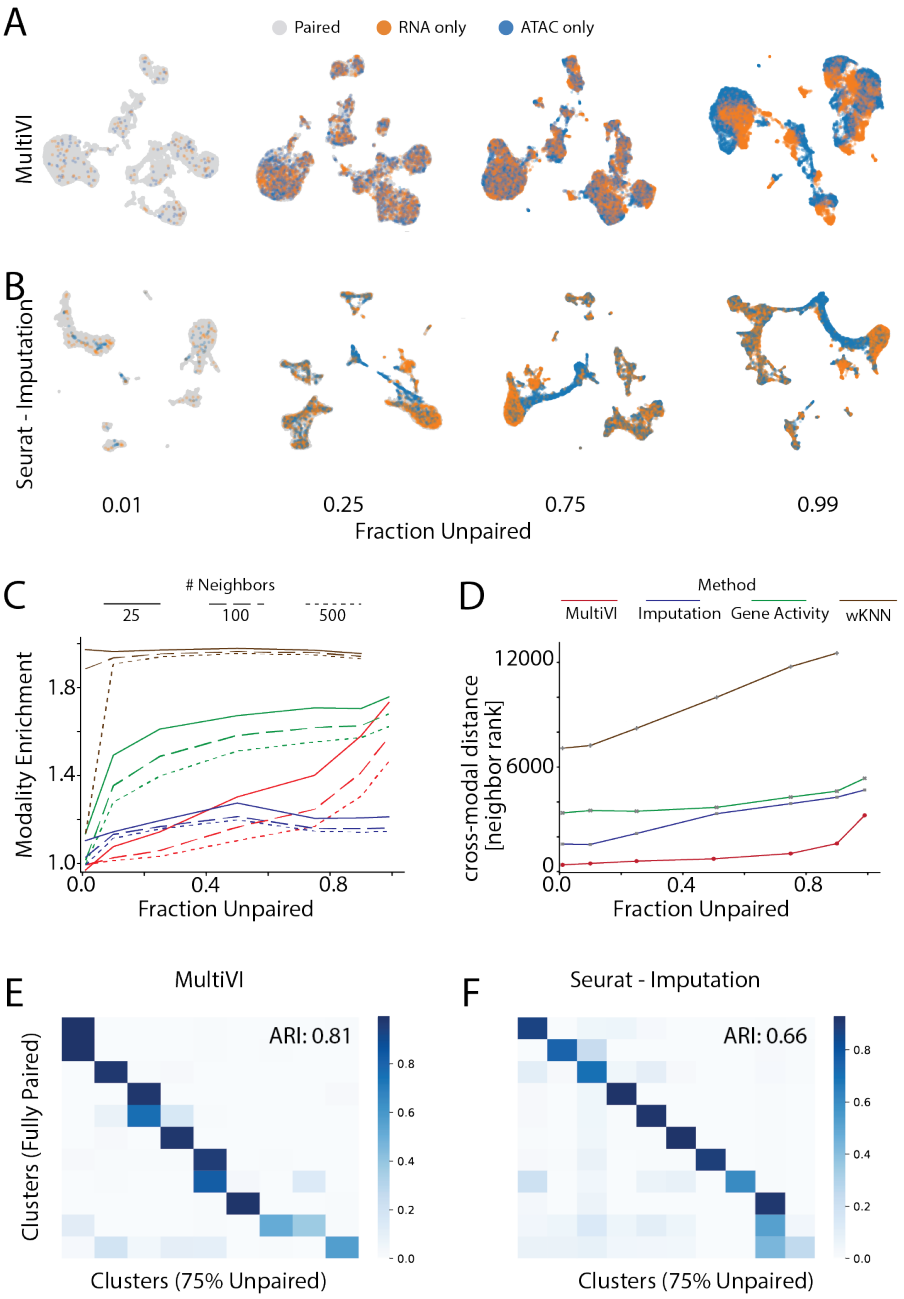

Figure 2: MultiVI accurately integrates gene expression and chromatin accessibility data. A) UMAP representations of the latent spaces learned by MultiVI (top row), Seurat using the RNA-only based integration (middle row), and Seurat using the WNN-based integration (bottom row) for various rates of unpaired data, colored by cell modality. B) The mean distance between the two representations of artificially unpaired cells, measured as the number of cells between them. C) Enrichment of neighbors from the same modality, computed as the fraction of neighbors of the K-nearest neighbors that are from the same modality, normalized by the overall fraction of cells from that modality. D-E) Correspondence between clusters identified using the latent space of MultiVI (D) and Seurat Imputation (E) methods run on the fully paired dataset versus clusters computed using the 0.75 fraction of unpaired cells data set.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

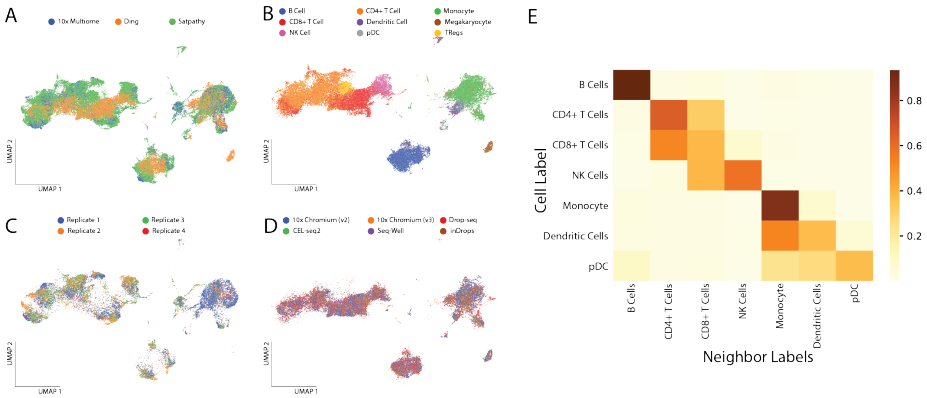

Figure 3: MultiVI effectively coalesce single and multi-modal data sets, integrating information from different experimental technologies and batches. We integrated three PBMC datasets in which only multi-modal data (10X multiome), only ATAC-seq information (Satpathy et al) and only RNA-seq information (Ding et al) is present. A)-D) UMAP representation computed from the latent space of MultiVI in which cells are color labeled by their dataset (A), their cell type (B) or ATAC-seq cells are labelled by the replicate in which they were collected (C) or RNA-seq cells are labelled by their collection experimental technology. E) Heatmap summarizing the fraction of cell neighbors that share the same label as the original cell.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

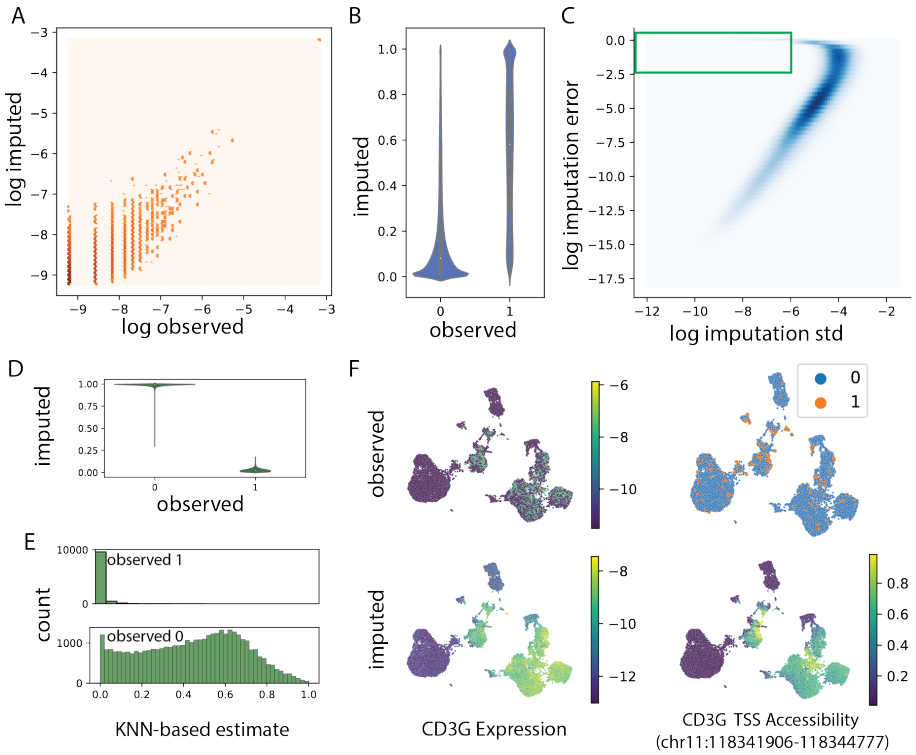

Figure 4: MultiVI imputation and uncertainty estimation. A) normalized observed RNA counts by MultiVI-imputed RNA estimates; all values, including color intensity, are presented in log scale ($log(x + 1e - 4)$ for stability). B) MultiVI-imputed accessibility estimates by the observed values. C) the imputation error $(\text{imputed} - \text{observed})^2$ as a function of the standard deviation of the imputed accessibility estimates. Green box marks high-confidence-high-error values examined in following panels. D) MultiVI-imputed accessibility estimates by the observed values for high-confidence-high-error cases. E) smooth accessibility estimates for values observed as 1 (top) and 0 (bottom). estimates computed by averaging the accessibility profiles of the 50 nearest neighbors, in a 50-dimensional space computed using Latent Semantic Indexing. F) observed and imputed values for CD3G expression and CD3G TSS accessibility. Expression values are normalized and in log scale.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

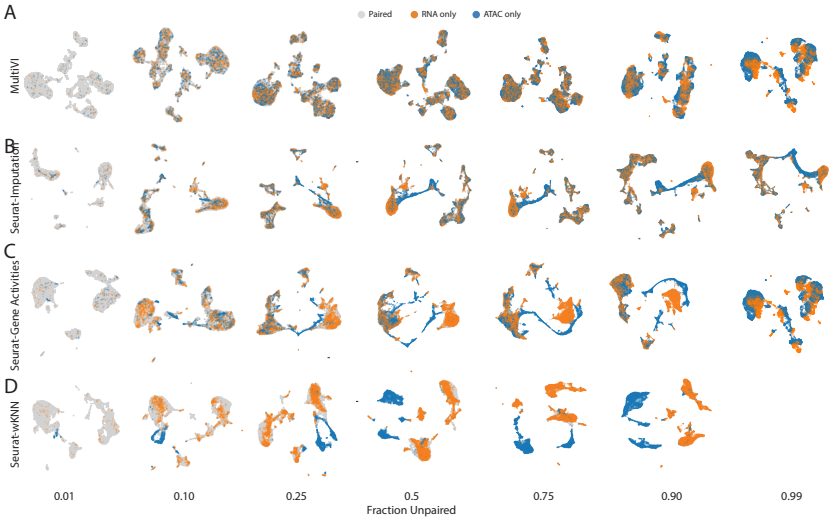

Figure S1: Extended Integration results depicting mixing of cells in data sets with different fraction of cells unpaired. UMAPS of latent representations for MultiVI (A), Seurat imputation method (B), Seurat Gene Activity Scores method (C), and Seurat wKNN method (D).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

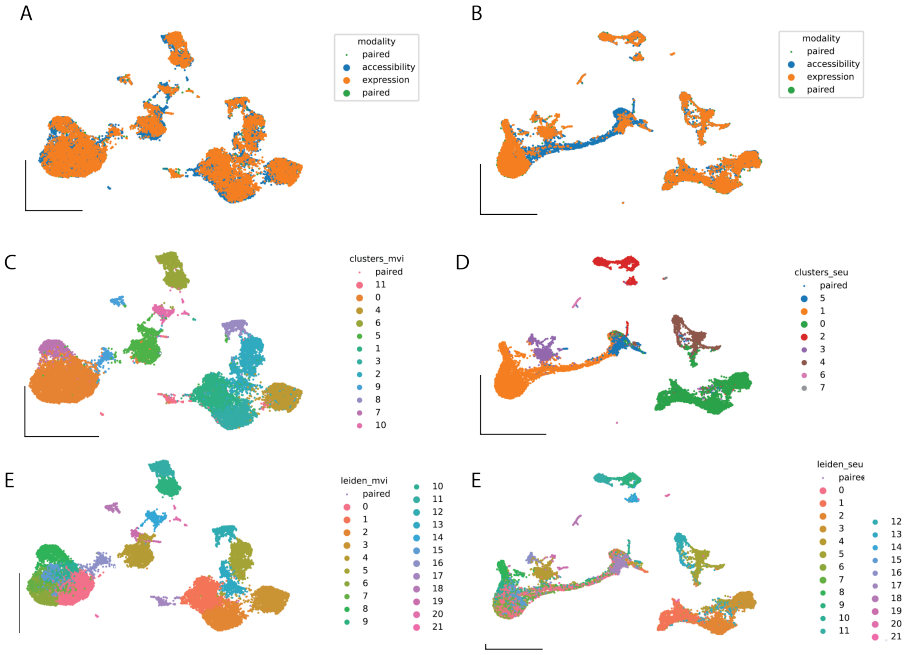

Figure S2: Extended Analysis of cluster consistency using dataset with 0.75 fraction of unpaired cells. UMAP representation computed from the latent space of MultiVI or Seurat Imputation in which cells are color labeled by their modality (A / B), their cluster correspondence computed at 0 fraction of unpaired cells (C / D) and their cluster correspondence computed at 0.75 fraction of unpaired cells (E / F).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

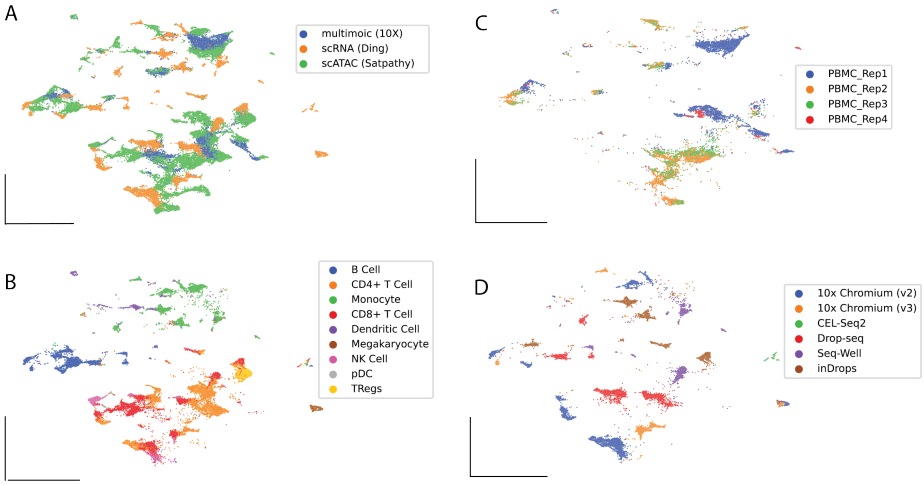

Figure S3: Latent representation of mixed sources data sets in which no batch correction techniques have been applied. We integrated three PBMC datasets in which only multi-modal data (10X multiome), only ATAC-seq information (Satpathy et al) and only RNA-seq information (Ding et al) is present without correcting for batch or modalities effects. A)-D) UMAP representation computed from the latent space of MultiVI in which cells are color labeled by their dataset (A), their cell type (B) or ATAC-seq cells are labelled by the replicate in which they were collected (C) or RNA-seq cells are labelled by their collection experimental technology.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.20.457057; this version posted August 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

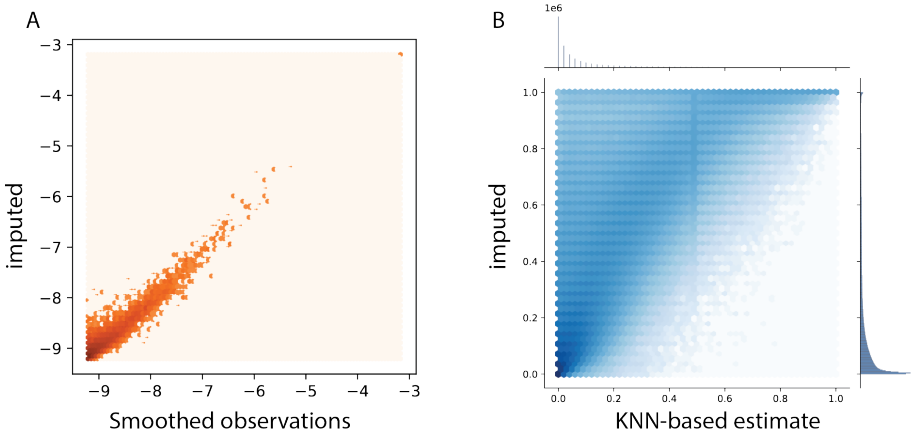

Figure S4: Imputed values compared with smoothed observations. Smooth averages of highly-similar cells (using 50 nearest neighbors in an independent low-dimensional space, computed separately for RNA and ATAC data) plotted against the MultiVI-imputed values for expression (A) and accessibility (B).