Benchmarking an autoencoder for integrating gene expression and chromatin accessibility modalities.

Katherine Feldmann[1], Manu Setty, PhD[2]

[1]Molecular and Cellular Biology – University of Washington, [2]Fred Hutchinson Cancer Center

Introduction

Single-cell sequencing technologies allow quantification of cellular heterogeneity, but sequencing individual modalities describes only one layer of the processes that regulate cellular function[1]. To determine a more comprehensive picture of cellular function, single-cells can be sequenced as paired modalities – for example, obtaining chromatin accessibility peaks and gene expression counts for individual cells. However, experimentally generating paired modalities is limited by complexity, cost and noise[2].

To overcome the limitations of experimentally generating multi-modal data sets, machine learning approaches have been developed to computationally derive paired modalities[3]. By training on small multi-modal data sets – which are more feasible to obtain – these cross-modality approaches learn the biological relationships between paired modalities rather than incorporating prior assumptions[4]. However, the low-dimensional latent space that represents these biological relationships is not interpretable[5]. Despite the potential applications for computationally pairing single-modal data sets, cross-modality approaches are not adequately benchmarked due to their recent innovation.

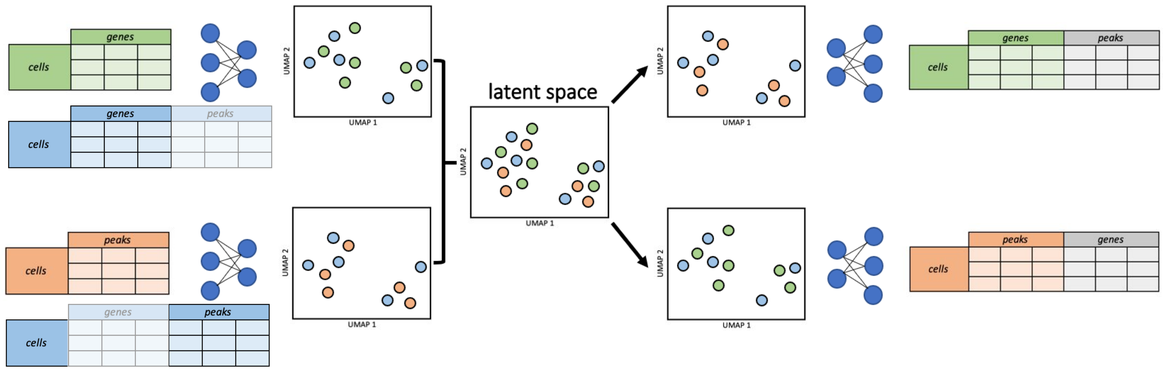


Figure 1 MultiVI is a variational autoencoder that uses neural networks to pair gene expression and chromatin accessibility modalities[6]. Modality-specific encoders learn biological relationships from the paired modalities, and modality-specific decoders sample from the low-dimensional latent space to integrate single-modalities or impute missing modalities.

How successful is MultiVI at integrating artificially unpaired gene expression and chromatin accessibility data?

Methods

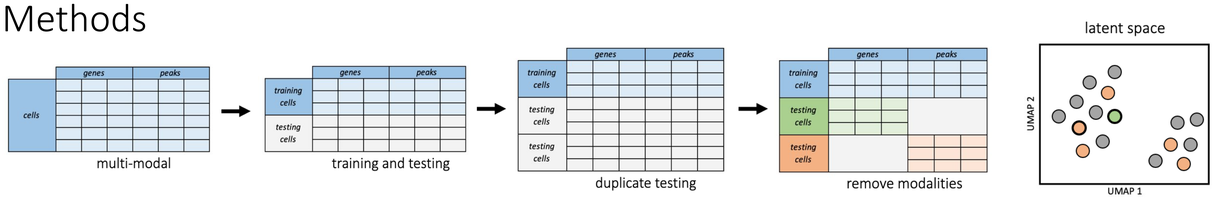


Figure 2 Duplicating and removing modalities from cells in a multi-modal data set creates single-modalities with ground truth matches that can be identified in the latent space.

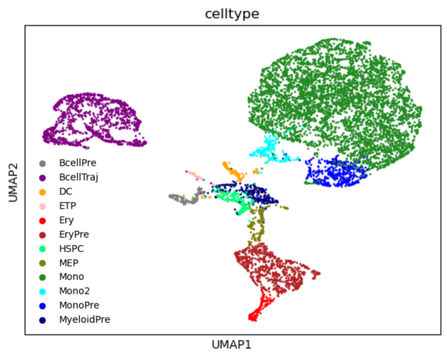


Figure 3 Cell types in the T-cell depleted multi-modal data set.

Dataset: 8,627 T-cell depleted bone marrow cells with paired gene expression counts and chromatin accessibility peaks.

Approach

1. Create ground truth matches by duplicating single-modality cells.
2. Determine the proximity of ground truth matches (and similar cells) in the latent space.
3. Compare the performance of single-modalities by integrating:
   gene expression ➔ chromatin accessibility
   chromatin accessibility ➔ gene expression

Results

Does MultiVI accurately match single-modal gene expression and chromatin accessibility data?

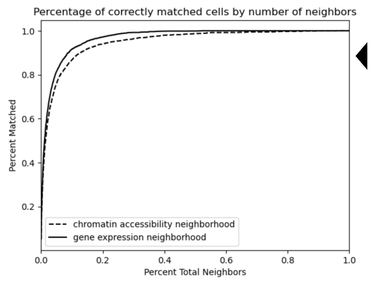


Figure 4 Paired modalities are retained in 70% of cells.

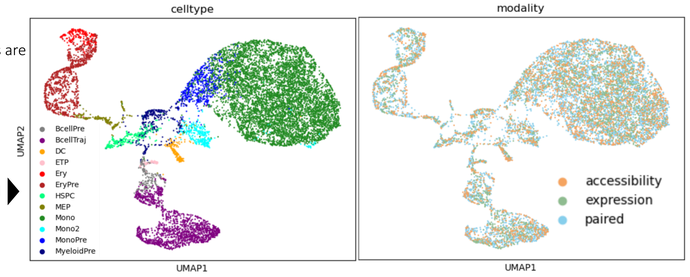


Figure 5 Modalities are evenly distributed in the low-dimensional latent space and cell types cluster together.

Gene expression neighborhood is better at integrating modalities than the chromatin accessibility neighborhood.

Does changing the proportion of cells with paired information affect the integration ability of unpaired cells?

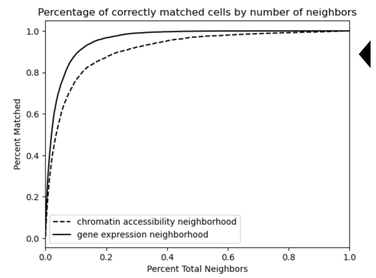


Figure 6 Paired modalities are retained in 30% of cells.

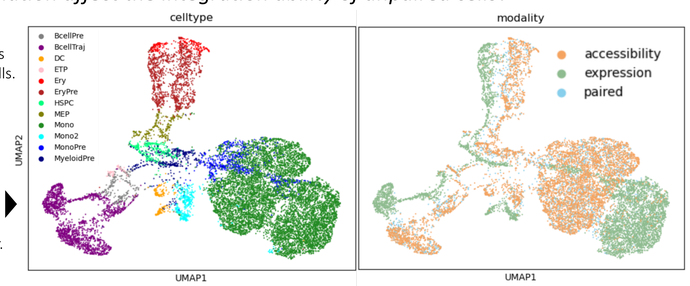
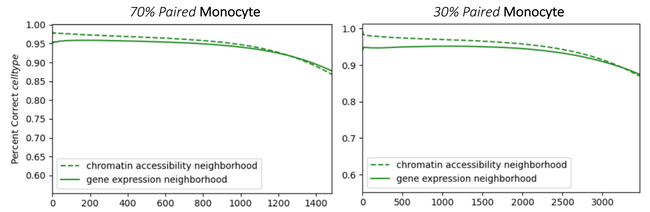


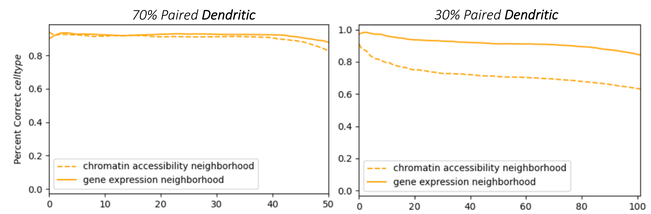
Figure 7 Modalities cluster separately in the low-dimensional latent space and cell types cluster together.

Limiting paired information causes a greater decrease in performance for the chromatin accessibility neighborhood.

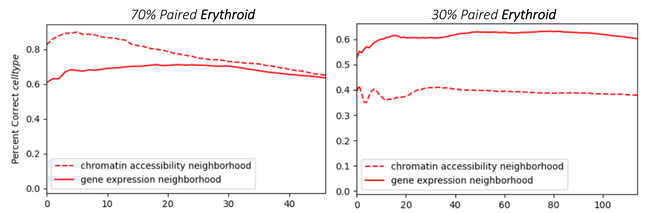
Are similar cells integrated if we do not sample the ground truth?



Majority of nearest neighbors are the correct cell type and limiting paired information has little effect.



Limiting paired information causes a greater decrease in chromatin accessibility neighborhood performance.



Percentage of nearest neighbors with the correct cell type is more variable for cell types with few cells.

Figure 8 Percent correct cell type by number of nearest neighbors. Neighbors are restricted by size of cell type.

Conclusions

MultiVI is better at integrating single-modalities when using the gene expression neighborhood, potentially because expression counts are more variable and provide more information than accessibility peaks. As expected with machine learning models, MultiVI performed better when trained with more paired information. Finally, performance varies by cell type where cell types with many cells are likely to integrate similar cells.

Acknowledgements

I would like to thank the members of the Setty Lab at Fred Hutchinson Cancer Center for their assistance with this project.

References

[1]Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, et al. 2023. Best practices for single-cell analysis across modalities. Nature Reviews Genetics.
[2]Wu KE, Yost KE, Chang HY, Zou J. 2021. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. PNAS.
[3]Lance C, Luecken MD, Burkhardt DB, Cannoodt R, Rautenstrauch P, Laddach A, et al. 2022. Multimodal single cell data integration challenge: results and lessons learned. bioRxiv.
[4,5]Hao Y, Stuart T, Kowalski M, Choudhary S, Hoffman P, Hartman A, et al. 2022. Dictionary learning for integrative, multimodal, and scalable single-cell analysis. bioRxiv.
[6]Ashuach T, Gabitto MI, Jordan MI, Yosef N. 2021. MultiVI: deep generative model for the integration of multi-modal data. bioRxiv.