Manuscript Title

This manuscript (permalink) was automatically generated from codingpoppy/multiomics_review_manubot@9dd643e on April 2, 2022.

Authors

- **John Doe**
  XXXX-XXXX-XXXX-XXXX · johndoe · johndoe
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Jane Roe**
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

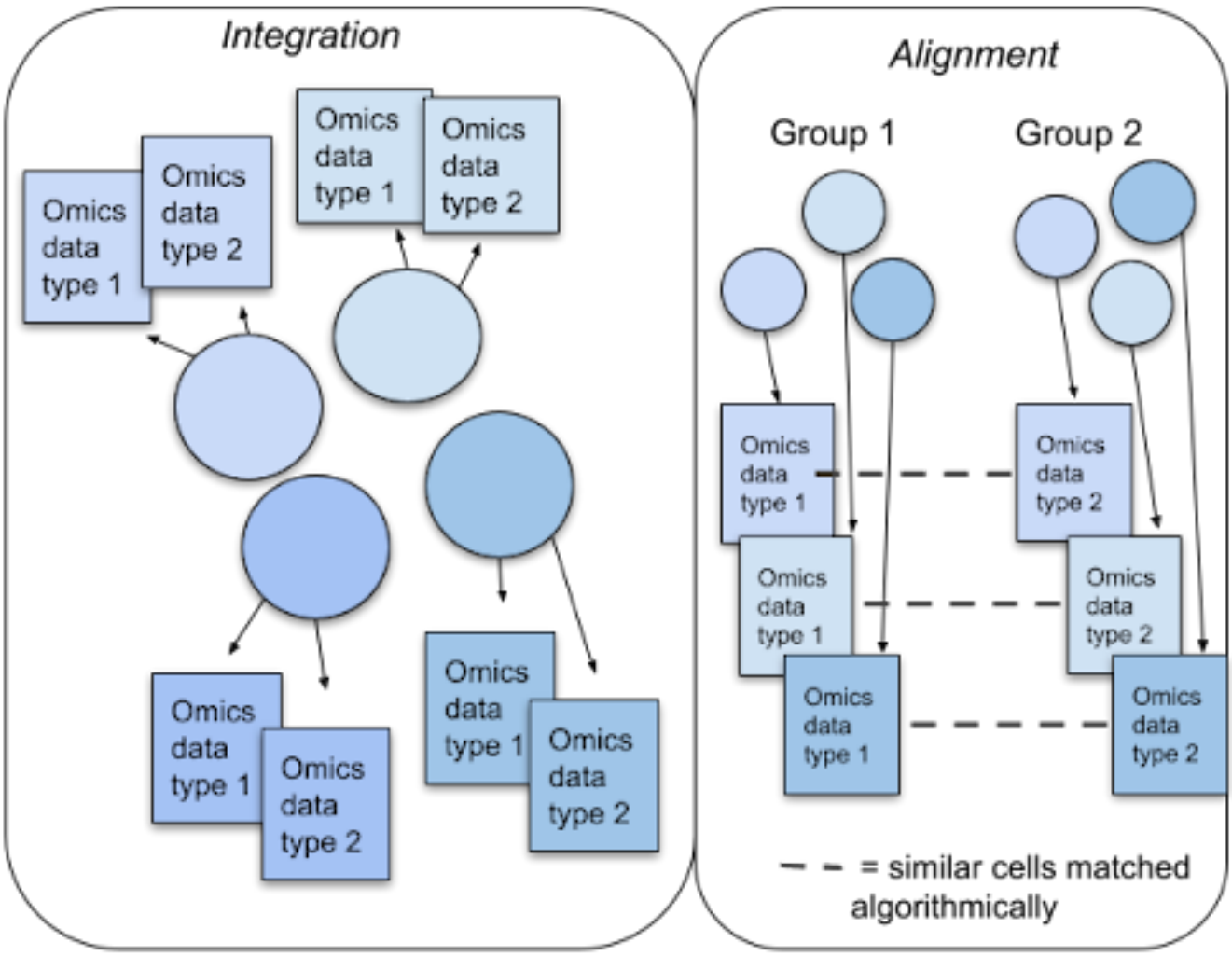
Abstract

Recently developed technologies to generate single-cell genomic data have made a revolutionary impact in the field of biology. Multi-omics assays offer even greater opportunities to understand cellular states and biological processes. However, the problem of integrating different -omics data with very different dimensionality and statistical properties remains quite challenging. A growing body of computational tools are being developed for this task, leveraging ideas ranging from machine translation to the theory of networks and representing a new frontier on the interface of biology and data science. Our goal in this review paper is to provide a comprehensive, up-to-date survey of computational techniques for the integration of multi-omics and alignment of multiple modalities of genomics data in the single cell research field.

Introduction

Single-cell sequencing technologies have opened the door to investigating biological processes at an unprecedentedly high resolution. Techniques such as DROP-seq [1] and 10x Genomics assays are capable of measuring single-cell gene expression, or scRNA-seq, in tens of thousands of single cells simultaneously. Measurements of other data modalities are also increasingly available. For example, single-cell ATAC-seq (scATAC-seq) assesses chromatin accessibility, and single-cell bisulfite sequencing captures DNA methylation, all from single cells. However, many of such techniques are designed to measure a single modality and do not lend themselves to multi-omics measurements. The way to combine information from such measurements is then to assay different -omics from different subsets of the same samples. By assuming that cells assayed by different techniques share similar properties, one can then use alignment methods to computationally aggregate similar cells across different omics assays and draw consensus biological inference.

Recently, however, a number of experimental techniques capable of assaying multiple modalities simultaneously from the same set of single cells have been developed. CITE-seq [2] and REAP-seq [3] measure proteins and gene expression. SNARE-seq [3,4], SHARE-seq [5] and sci-CAR [6] measure gene expression and chromatin accessibility, while scGEM [**doi?** 10.1038/nmeth.3961] measures gene expression and DNA methylation. For triple-omics data generation, scNMT [7] measures gene expression, chromatin accessibility and DNA methylation, and scTrio-seq [8,9] captures SNPs, gene expression and DNA methylation simultaneously. Integrative analysis of such data obtained from the same cells remains a challenging computational task due to a combination of reasons, such as the noise and sparsity in the assays, and different statistical distributions for different modalities. For clarity, we distinguish between integration methods that combine multiple -omics data from the set of the same single cells (Section I), from alignment methods designed to work with multi-modal data coming from the same tissue but different cells (Section II). The difference in their approaches is shown in Figure. {[**fig-1?**]}.



Multi-omics data can sometimes be sequenced from the same set of single cells (left); at other times, only the data sequenced from the same/similar sample, but different single cells are available (right). In the former case, we have the task of integrating the different data modalities (left); in the latter case, we need to first identify similar cells across the samples (right) - this is the computational task of alignment.

The application of data fusion algorithms for multi-omics sequencing data predates the single-cell technologies; bulk-level data have been integrated using a variety of computational tools as reviewed in [10]. In this review, we aim to give a comprehensive, up-to-date summary of existing computational tools of multi-omics data integration and alignment in the single-cell field, for researchers in the field of computational biology. For more general surveys, the readers are encouraged to check other single-cell multi-omics reviews [11,12,13,14,15,16].

Integration methods handling multi-omics data generated from the same single cells

The integration methods for multi-modal data assayed from the same set of single cells can be broadly categorized into at least three main types by methodology: mathematical matrix factorization methods, AI (eg. neural-network) based methods and network-based methods. The scheme of these methods is illustrated in Figure {[**fig-2?**]}. Additional less diversified approaches include a Bayesian statistical method and a metric learning method. The list of the currently implemented methods is summarized in Table [**tbl-1?**].

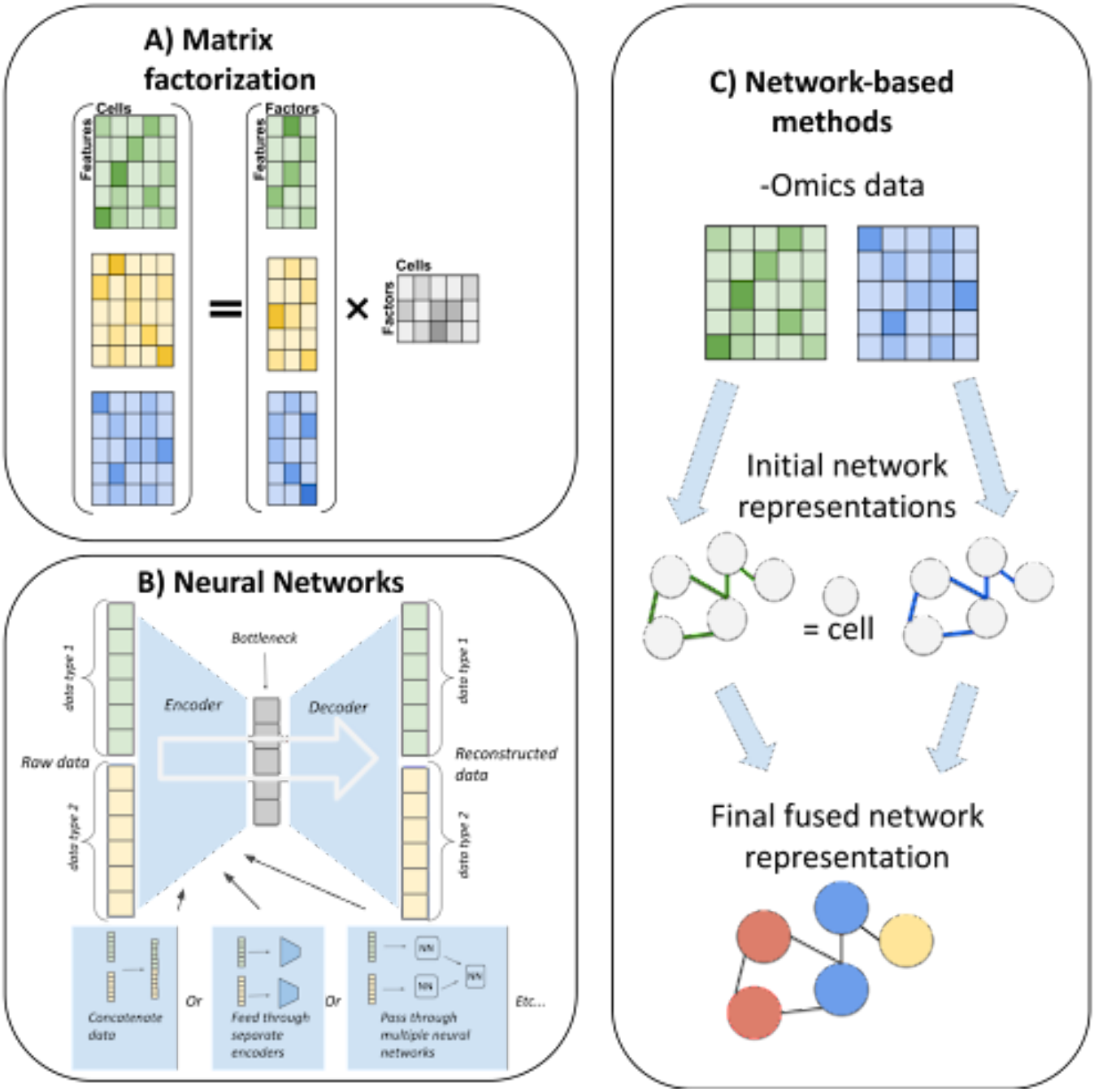


Illustration of some common integration approaches for single-cell multi-omics: matrix factorization, neural network and network-based approaches.

Methodology Category	Method	Data	Algorithm	Reference
Matrix Factorization	MOFA+	Transcriptomic, Epigenetic	Matrix Factorization with Automatic Relevance Determination	[8]
	scAI	Transcriptomic, Epigenetic	Matrix factorization, with custom aggregation of epigenetic data	[10]
Neural Network	totalVI	Transcriptomic, Proteomic	Variational autoencoder	[12]
	scMVAE	Transcriptomic, Epigenetic		[13]
	DCCA	Transcriptomic, Epigenetic		[17]

Methodology Category	Method	Data	Algorithm	Reference
	LIBRA	Transcriptomic, Proteomic, Epigenetic	Split-brain autoencoder	[16]
	BABEL	Transcriptomic, Proteomic, Epigenetic	Autoencoder translating between modalities	[18]
	DeepMAPS	Transcriptomic, Epigenetic, Proteomic	Graph Neural Network	[19]
Network - Based	citeFUSE	Transcriptomic, Proteomic	Similarity network fusion	[20]
	Seurat v4	Transcriptomic, Proteomic	Weighted averaging of nearest neighbor graphs	[21]
	Integrated Diffusion	Transcriptomic	Joint Manifold Learning through Integrated Diffusion	[22]
Other	BREM-SC	Transcriptomic, Proteomic	Bayesian mixture model	[23]
	SCHEMA	Transcriptomic, Epigenetic	Metric Learning	[24]

Table {#tbl-1}: Summary of the methods for integrating multi-omics data from the same cells.

Matrix Factorization based methods

Matrix factorization methods aim to describe each cell as the product between a vector that describes each -omics element (genes, epigenetic loci, proteins, etc.) and a vector of reduced and common features (“factors”) capturing its basic properties (Figure {[fig-2?]}A). Mathematically, if we represent each -omics as matrix

$$X_{i(i=1,2,\cdots)}$$

then matrix factorization decomposes it as the product of a shared matrix H across all omics data types, and -omics specific matrix

$$W_{i(i=1,2,\cdots)}$$

, together with random noise

$$\epsilon_{i(i=1,2,\cdots)}$$

as

$$X_1 = W_1 H + \epsilon_1, X_2 = W_2 H + \epsilon_2, \cdots, X_i = W_i H + \epsilon_i$$

Such methods are simple and easily interpretable since the cell and -omics factors both carry clearly discernible biological meaning, but may lack the ability to capture nonlinear effects. We describe the variations in this type of methods below:

MOFA+ [25] is a sequel to the MOFA (Multi-Omics Factor Analysis) [20]. Both studies perform factor analysis, equipped with sparsity-inducing Bayesian elements including Automatic Relevance Determination [26]. MOFA+ integrates data over both views (corresponding to different modalities) and groups (corresponding to different experimental conditions). The model scales easily to large datasets. MOFA+ was applied to integrate gene expression, chromatin accessibility and DNA methylation data assayed using scNMT from mouse embryos, as well as to integrate several datasets over different experimental conditions rather than different -omics. After performing factor analysis on the mouse dataset, the most relevant factors are related to biological processes shaping embryo development. MOFA+ provides an elegant and successful general framework for integration, which could potentially be superseded in specific cases by more specialized models designed for integrating specific -omics layers.

scAI (“single-cell aggregation and inference”) [doi:10.1186/s13059-020-1932-8] features a twist on matrix factorization and is designed specifically for integration of epigenetic (chromatin accessibility, DNA methylation) and transcriptomic data. It addresses the sparsity of epigenetic data by aggregating (averaging) such data between similar cells. This requires a notion of cell-cell similarity which is learned as a part of the model, rather than being postulated prior to the integration. Their model solves the following optimization problem

$$\min_{W_1,W_2,H,Z} \alpha||X_1 - W_1 H||_F^2 + ||X_2(Z \cdot R - W_2 H)||_F^2 + \lambda||Z_H^T H||_F^2 + \gamma \sum_j ||H_{\cdot j}||_1^2$$

where

X_1

represents the transcriptomic data,

X_2

the epigenomic data,

H

are the common (cell-specific) factors,

W_1

,

W_2

are the assay-specific factors,

Z

is the cell-cell similarity matrix, and entries of are Bernoulli-distributed random variables. The twist on the usual matrix factorization is made by factoring aggregated epigenetic data

$X_2(Z \cdot R)$

, rather than directly factoring the epigenetic data

X_2

. After the learning is complete, the matrix of cell factors is used to cluster the cells and the importance of genes and epigenetic marks is ranked using the magnitude of the values in loading matrices. In order to jointly visualize different factors, scAI implements a novel VscAI algorithm utilizing Sammon mappings [27]. The relationships between epigenetics and gene expression can be explored using correlation analysis and nonnegative least square regression. The model was tested on simulations using MOSim [28], and several real world datasets, and performed better than the earlier MOFA version, in terms of identifying natural clusters and condensing epigenetic data into meaningful factors.

Neural Network based methods

While neural networks are generally well-suited for supervised tasks, a class of neural networks called autoencoders is commonly used for unsupervised learning, such as the multi-omics integration problem in single cells. Deep autoencoders perform nonlinear dimensionality reduction by squeezing the input through a lower-dimensional hidden layer (“bottle neck”) and attempting to reconstruct the original input as the output of the neural network (Figure {[fig-2?]}B). They consist of two parts: the “encoder” network performing the dimensionality reduction and the “decoder” network reconstructing based on the dimensionally reduced data. In principle, autoencoders generalize the principal component analysis by allowing for nonlinear transformations. Many variations of autoencoder models exist, and among them variational autoencoders have proven useful for analyzing single-cell data. Rather than directly encoding the data in a dimensionally reduced (“latent”) space, variational autoencoders sample from a probability distribution (usually Gaussian) in the latent space, and use the encoder network to produce the parameters of this distribution. As such, they combine deep learning and Bayesian inference to produce generative models, which not only dimensionally reduce the original data but also produce realistic synthetic data points. Below we review the methods using certain variations of the autoencoder architecture to integrate single-cell multi-omics data.

scMVAE (“Single Cell Multimodal Variational Autoencoder”) [29] was designed to integrate transcriptomic and chromatin accessibility data, using a version of a variational autoencoder. The key question in multi-omics integration is how to encode the multi-omics data into a single latent space representation. In the case of scMVAE, a combination of 3 different methods was used for this task, including a neural network acting on the concatenated input data, neural networks encoding transcriptomic and chromatin accessibility data separately prior to merging, and a “Product of Experts” technique for combining different representations [30]. At the same time, cell-specific scales used to normalize expression across cells are learned (called “library factors”). The input data are reconstructed by processing the latent representations via decoder neural networks, which calculate the probabilities of gene dropouts and predict the expression of measured genes modelled as a negative binomial distribution.

This model incorporates the task of constructing shared representations of the multi-modal data with clustering. Namely, one of the latent variables is constructed to correspond to the clustering label c. Furthermore, the model incorporates tools to deal with tasks such as data imputation, and can be used for studying the association between epigenetics and gene expression. scMVAE was applied to integrate two real datasets assaying mRNA and chromatin accessibility using SNARE - seq method, as well as simulated data generated by “Splatter” [31]. It takes into account the known relationships between appropriately located transcription factors and gene expression, and uses them to test the imputed (denoised) data. According to the authors, scMVAE performed better than MOFA in terms of clustering and enhancing the consistency between different -omics layers on several real and simulated datasets.

DCCA, denoting “Deep cross-omics cycle attention model”, is another method in this category for joint analysis of single-cell multi-omics data [17]. It uses variational autoencoders to integrate multi-omics data, and builds on the scMVAE algorithm described above. However, DCCA diverges from scMVAE in one important aspect: DCCA uses separate but coupled autoencoders to dimensionally reduce different -omics layers, while scMVAE constructs a shared dimensionally reduced representation of transcriptomic and epigenetic data. This strategy is inspired by the theory of machine translation, notably the so-called “attention transfer”; in this case, the “teacher network” working with the scRNA-seq data guides the learning of the “student network” working with scATAC-seq data. Their model compares favorably to scAI and MOFA+ on metrics such as clustering accuracy, denoising quality and consistency between different -omics.

totalVI [32] combines Bayesian inference and a neural network to create a generative model for data integration. It was created to handle gene expression and protein data. Joint latent space representations are learned via an encoder network and used to reconstruct the original data while accounting for the difference between the original data modalities. The model generates latent representations capturing both -omics, and at the same time models experimental conditions through an additional set of latent variables. The gene expression data are sampled from a negative binomial distribution, and the parameters are obtained as outputs of a decoder neural network. The protein data are sampled from a mixture model with two negative binomial distributions simulating the experimental background and the actual signal respectively. The model was applied to two datasets containing transcriptomic and proteomic measurements, and generated shared representations of cells with interpretable components.

LIBRA [33] uses an autoencoder-like neural network to “translate” between different omics. Motivated by “split-brain autoencoder”[34], and “machine translation” approach, the model consists of two separate neural networks. The first network takes as input elements of the first dataset and aims to reconstruct a corresponding element of the second dataset. The second network performs an inverse task. Taken together, the bottlenecks of two networks aim to convert the two datasets into the same latent space. This method is quite general and can be applied to various pairs of -omics data. It produced clusters of similar quality compared to Seurat v4.

BABEL [18] also uses autoencoder-like neural networks to translate between gene expression (modeled by Negative Binomial distribution) and binarized chromatin accessibility data. There are two encoder and two decoder neural networks, each encoder/decoder handles one data type of gene expression or chromatin accessibility. As a result, four combinations between encoders and decoders are formed, and the loss function is optimized to minimize reconstruction error for four combinations of encoders and decoders. In this approach, the two encoders are prone to produce similar representations, as the encoded gene accessibility is decoded as chromatin accessibility and vice versa. BABEL provides a promising generic framework to multi-omics inference at a single-cell level from single-omics data, by using the model that was previously trained on multi-omics data sequenced from the same single cells. The modular nature of BABEL provides additional flexibility, as the model can be extended to work with additional modalities when the corresponding data becomes available. Despite the potential for generalization, one should be cautioned that if the training is conducted on cell types that are very different, the transfer learning using BABEL is not very successful.

DeepMAPS [19] integrates different data modalities by a graph transformer neural network architecture for interpretable representation learning. The data is represented using a heterogenous graph in which some of the nodes represent cells and others represent genes. An autoencoder-like graph neural network architecture is used for representation learning, with an attention mechanism. The attention mechanism learns the weights by the contribution of the neighbors to the node of interest. This not only achieves better performance, but also enhances the interpretability to identify genes most relevant to cell state differences. DeepMAPS method learns relevant gene-gene interaction networks and cell-cell similarities, which can be used for downstream steps such as clustering to infer novel cell types. It compared favorably on clustering, compared to state-of-the art techniques such as MOFA+ and totalVI.

Network-based methods

Network-based methods represent the relationships between different cells using a weighted graph, where cells serve as nodes (Figure {[**fig-2?**]}C). Integration is then accomplished by manipulating such graph representation. This approach emphasizes the neighborhood structure and sometimes pools the information between neighbors, leading to additional robustness against the noise. Below are the currently available methods.

citeFUSE [35] integrates transcriptomic and proteomic CITE-seq data using network fusion of similarity graphs corresponding to different modalities. This idea traces back to computer science work [36] on fusing multi-view networks through cross-diffusion, and to the follow-up SNF method [37] that was used to integrate bulk level multi-omics data. The algorithm adjusts the graph connectivities by a process of diffusion, which allows for the distance information to be aggregated between neighbors. Namely, the algorithm consists of two iterative steps: separate diffusion on different -omics layers and fusion across the -omics layers. It results in a fused consensus matrix of distances between cells, borrowing information from multiple -omics. citeFUSE used spectral clustering to identify cell types, and showed an improvement over single-modality based clusters. Additional benefits of the method include inference of ligand-receptor interactions and a novel tool for doublet detection.

Joint Diffusion [22] constructs graph representations of different -omics and then performs a joint diffusion process on the two graphs in order to denoise and integrate the data. This approach builds upon MAGIC [38], a method for denoising scRNA-seq data, and generalizes it to multi-modal data. Diffusion can be conceptualized as a random walk process. In a graph diffusion algorithm, random walking on the graph can help discover the intrinsic structure of the data hidden behind the noise. In Joint Diffusion random walks are performed while allowing for transitions from one graph to another. A key idea in this work is to quantify the amount of noise in different datasets, through a spectral

entropy of the corresponding graphs, and adjust the time one spends on different graphs in accordance with their relative levels of noise. In this way, the transcriptomic and epigenetic data will not be weighted equally, as the transcriptomic data is generally of better quality. This method excels at denoising and visualizations, and was shown to present an improved clustering performance compared to single-modality clustering and the one based on a more naive alternating diffusion process.

Seurat v4 [39] aims to represent the data as a WNN (weighted nearest neighbor) graph in which cells that are similar according to the consensus of both modalities are connected. In the process of constructing a WNN graph, a set of cell-specific weights dictating the relative importance of different -omics data is learned. Such weights often carry important biological meaning. Specifically, Seurat v4 pipeline has the following steps: first, data corresponding to different -omics are dimensionally reduced using PCA to the same number of dimensions. Then, kNN (k nearest neighbor) graphs corresponding to different -omics are constructed. In a kNN graph, each datapoint (a node of this graph) is connected to nearest neighboring nodes. Cell-specific coefficients determining the relative importance of different -omics are then learned by considering the accuracy of inter-modality and cross-modality predictions by nearest neighbor graphs. Lastly, a linear combination of data from different omics is done, using the coefficients learned in the previous step. The nearest neighbors with respect to those linear combinations are then connected to build the WNN graph. Seurat v4 was applied to a CITE-seq based transcriptomic and proteomic dataset, and several other datasets involving mRNA, proteins and chromatin accessibility. The authors compared this method with MOFA+ and totalVI, using correlations (Pearson and Spearman) between the data corresponding to a cell and the average of its nearest latent space neighbors, and claimed that it performed better than MOFA+ or totalVI.

Other Models

BREMSC [23] is a Bayesian mixture method. It integrates single-cell gene expression and protein data by modeling them as a mixture of probability distributions that share the same underlying set of parameters. The model is useful for performing joint clustering, where confidence in cluster assignments can be quantified via posterior probabilities. It performed favorably compared to single-omics clustering methods. While the MCMC procedure used to train the model can be computationally intensive, the model provides an effective way of integration by accounting the differences between the two -omics layers using probability distributions.

SCHEMA [24] is a different metric learning approach that aims to construct a notion of distances on the space of samples, taking into account different -omics data. One of the -omics (usually, scRNA-seq) is considered the primary base for distance, additional omics are then used to modify this distance. This is formulated as optimization of the quadratic function using quadratic programming. The scRNA-seq and scATAC-seq data can thus be integrated, yielding downstream insights into cell developmental trajectories. This method showed a better clustering performance than those based on clustering different modalities separately or integrating them using canonical correlation analysis. It is a useful method for asymmetrically integrating data modalities of different qualities, such as the case of scRNA-seq and scATAC-seq data.

References

1. **Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets**
Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, … Steven A McCarroll
Cell (2015-05) https://doi.org/f7dkxv
DOI: 10.1016/j.cell.2015.05.002 · PMID: 26000488 · PMCID: PMC4481139

2. **Simultaneous epitope and transcriptome measurement in single cells**
Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, Peter Smibert
Nature Methods (2017-07-31) https://doi.org/gfkksd
DOI: 10.1038/nmeth.4380 · PMID: 28759029 · PMCID: PMC5669064

3. **Multiplexed quantification of proteins and transcripts in single cells**
Vanessa M Peterson, Kelvin Xi Zhang, Namit Kumar, Jerelyn Wong, Lixia Li, Douglas C Wilson, Renee Moore, Terrill K McClanahan, Svetlana Sadekova, Joel A Klappenbach
Nature Biotechnology (2017-08-30) https://doi.org/gghhzg
DOI: 10.1038/nbt.3973 · PMID: 28854175

4. **High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell**
Song Chen, Blue B Lake, Kun Zhang
Nature Biotechnology (2019-10-14) https://doi.org/gghfzh
DOI: 10.1038/s41587-019-0290-0 · PMID: 31611697 · PMCID: PMC6893138

5. **SHARE-seq reveals chromatin potential**
Dorothy Clyde
Nature Reviews Genetics (2020-11-17) https://doi.org/gps358
DOI: 10.1038/s41576-020-00308-6 · PMID: 33204030

6. **Joint profiling of chromatin accessibility and gene expression in thousands of single cells**
Junyue Cao, Darren A Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A Pliner, Andrew J Hill, Riza M Daza, Jose L McFaline-Figueroa, Jonathan S Packer, Lena Christiansen, … Jay Shendure
Science (2018-09-28) https://doi.org/gd4rk4
DOI: 10.1126/science.aau0730 · PMID: 30166440 · PMCID: PMC6571013

7. **scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells**
Stephen J Clark, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M Stubbs, Heather J Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C Marioni, … Wolf Reik
Nature Communications (2018-02-22) https://doi.org/gc4q72
DOI: 10.1038/s41467-018-03149-4 · PMID: 29472610 · PMCID: PMC5823944

8. **Single-cell multimodal profiling reveals cellular epigenetic heterogeneity**
Lih Feng Cheow, Elise T Courtois, Yuliana Tan, Ramya Viswanathan, Qiaorui Xing, Rui Zhen Tan, Daniel SW Tan, Paul Robson, Yuin-Han Loh, Stephen R Quake, William F Burkholder
Nature Methods (2016-08-15) https://doi.org/gps357
DOI: 10.1038/nmeth.3961 · PMID: 27525975

9. **Single-cell multiomics sequencing and analyses of human colorectal cancer**

Shuhui Bian, Yu Hou, Xin Zhou, Xianlong Li, Jun Yong, Yicheng Wang, Wendong Wang, Jia Yan, Boqiang Hu, Hongshan Guo, ... Wei Fu
Science (2018-11-30) https://doi.org/gfkwwk
DOI: 10.1126/science.aao3791 · PMID: 30498128

10. **More Is Better: Recent Progress in Multi-Omics Data Integration Methods**
Sijia Huang, Kumardeep Chaudhary, Lana X Garmire
Frontiers in Genetics (2017-06-16) https://doi.org/gcz6m3
DOI: 10.3389/fgene.2017.00084 · PMID: 28670325 · PMCID: PMC5472696

11. **Statistical single cell multi-omics integration**
M Colomé-Tatché, FJ Theis
Current Opinion in Systems Biology (2018-02) https://doi.org/gfgkr6
DOI: 10.1016/j.coisb.2018.01.003

12. **Integrative Methods and Practical Challenges for Single-Cell Multi-omics**
Anjun Ma, Adam McDermaid, Jennifer Xu, Yuzhou Chang, Qin Ma
Trends in Biotechnology (2020-09) https://doi.org/ggzm72
DOI: 10.1016/j.tibtech.2020.02.013 · PMID: 32818441 · PMCID: PMC7442857

13. **Computational methods for the integrative analysis of single-cell data**
Mattia Forcato, Oriana Romano, Silvio Bicciato
Briefings in Bioinformatics (2020-08-06) https://doi.org/gpfxqp
DOI: 10.1093/bib/bbaa042 · PMID: 32363378 · PMCID: PMC7820847

14. **Computational principles and challenges in single-cell data integration**
Ricard Argelaguet, Anna SE Cuomo, Oliver Stegle, John C Marioni
Nature Biotechnology (2021-05-03) https://doi.org/gjw92q
DOI: 10.1038/s41587-021-00895-7 · PMID: 33941931

15. **Computational strategies for single-cell multi-omics integration**
Nigatu Adossa, Sofia Khan, Kalle T Rytkönen, Laura L Elo
Computational and Structural Biotechnology Journal (2021) https://doi.org/gmh28b
DOI: 10.1016/j.csbj.2021.04.060 · PMID: 34025945 · PMCID: PMC8114078

16. **Multi-omics integration in the age of million single-cell data**
Zhen Miao, Benjamin D Humphreys, Andrew P McMahon, Junhyong Kim
Nature Reviews Nephrology (2021-08-20) https://doi.org/gmzwc3
DOI: 10.1038/s41581-021-00463-x · PMID: 34417589

17. **Deep cross-omics cycle attention model for joint analysis of single-cell multi-omics data**
Chunman Zuo, Hao Dai, Luonan Chen
Bioinformatics (2021-05-24) https://doi.org/gj5qs3
DOI: 10.1093/bioinformatics/btab403 · PMID: 34028557

18. **BABEL enables cross-modality translation between multi-omic profiles at single-cell resolution**
Kevin E Wu, Kathryn E Yost, Howard Y Chang, James Zou
Cold Spring Harbor Laboratory (2020-11-10) https://doi.org/gps5d7
DOI: 10.1101/2020.11.09.375550

19. **DeepMAPS: Single-cell biological network inference using heterogeneous graph transformer**
Anjun Ma, Xiaoying Wang, Cankun Wang, Jingxian Li, Tong Xiao, Juexing Wang, Yang Li, Yuntao Liu, Yuzhou Chang, Duolin Wang, ... Qin Ma
Cold Spring Harbor Laboratory (2021-11-03) https://doi.org/gps5d9

DOI: 10.1101/2021.10.31.466658

20. **Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets**
Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, Oliver Stegle
Molecular Systems Biology (2018-06) https://doi.org/gdqq3f
DOI: 10.15252/msb.20178124 · PMID: 29925568 · PMCID: PMC6010767

21. **scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles**
Suoqin Jin, Lihua Zhang, Qing Nie
Genome Biology (2020-02-03) https://doi.org/gps5fb
DOI: 10.1186/s13059-020-1932-8 · PMID: 32014031 · PMCID: PMC6996200

22. **Multimodal Data Visualization and Denoising with Integrated Diffusion**
Manik Kuchroo, Abhinav Godavarthi, Alexander Tong, Guy Wolf, Smita Krishnaswamy
arXiv (2021) https://doi.org/gps5fc
DOI: 10.48550/arxiv.2102.06757

23. **BREM-SC: a bayesian random effects mixture model for joint clustering single cell multi-omics data**
Xinjun Wang, Zhe Sun, Yanfu Zhang, Zhongli Xu, Hongyi Xin, Heng Huang, Richard H Duerr, Kong Chen, Ying Ding, Wei Chen
Nucleic Acids Research (2020-05-07) https://doi.org/gm37hd
DOI: 10.1093/nar/gkaa314 · PMID: 32379315 · PMCID: PMC7293045

24. **Schema: metric learning enables interpretable synthesis of heterogeneous single-cell modalities**
Rohit Singh, Brian L Hie, Ashwin Narayan, Bonnie Berger
Genome Biology (2021-05-03) https://doi.org/gjzh4x
DOI: 10.1186/s13059-021-02313-2 · PMID: 33941239 · PMCID: PMC8091541

25. **MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data**
Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, Oliver Stegle
Genome Biology (2020-05-11) https://doi.org/ggvwsr
DOI: 10.1186/s13059-020-02015-1 · PMID: 32393329 · PMCID: PMC7212577

26. **Bayesian learning for neural networks**
Radford M Neal
Springer (1996)
ISBN: 9780387947242

27. **A Nonlinear Mapping for Data Structure Analysis**
JW Sammon
IEEE Transactions on Computers (1969-05) https://doi.org/frbptp
DOI: 10.1109/t-c.1969.222678

28. **MOSim: Multi-Omics Simulation in R**
Carlos Martínez-Mira, Ana Conesa, Sonia Tarazona
Cold Spring Harbor Laboratory (2018-09-20) https://doi.org/gkwhng
DOI: 10.1101/421834

29. **Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data**
Chunman Zuo, Luonan Chen
Briefings in Bioinformatics (2020-11-17) https://doi.org/gps5d4
DOI: 10.1093/bib/bbaa287 · PMID: 33200787 · PMCID: PMC8293818

30. **Training Products of Experts by Minimizing Contrastive Divergence**
Geoffrey E Hinton
Neural Computation (2002-08-01) https://doi.org/ct8dwx
DOI: 10.1162/089976602760128018 · PMID: 12180402

31. **Splatter: simulation of single-cell RNA sequencing data**
Luke Zappia, Belinda Phipson, Alicia Oshlack
Genome Biology (2017-09-12) https://doi.org/gc3h3g
DOI: 10.1186/s13059-017-1305-0 · PMID: 28899397 · PMCID: PMC5596896

32. **A Joint Model of RNA Expression and Surface Protein Abundance in Single Cells**
Adam Gayoso, Romain Lopez, Zoë Steier, Jeffrey Regier, Aaron Streets, Nir Yosef
Cold Spring Harbor Laboratory (2019-10-07) https://doi.org/dcnc
DOI: 10.1101/791947

33. **Machine Translation between paired Single Cell Multi Omics Data**
Xabier Martinez-de-Morentin, Sumeer A Khan, Robert Lehmann, Jesper Tegner, David Gomez-Cabrero
Cold Spring Harbor Laboratory (2021-01-28) https://doi.org/gps5d8
DOI: 10.1101/2021.01.27.428400

34. **Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction**
https://ieeexplore.ieee.org/document/8099559

35. **CiteFuse enables multi-modal analysis of CITE-seq data**
Hani Jieun Kim, Yingxin Lin, Thomas A Geddes, Jean Yang, Pengyi Yang
Cold Spring Harbor Laboratory (2019-11-25) https://doi.org/gktx8j
DOI: 10.1101/854299

36. **Unsupervised Metric Fusion Over Multiview Data by Graph Random Walk-Based Cross-View Diffusion** https://ieeexplore.ieee.org/document/7348699

37. **Similarity network fusion for aggregating data types on a genomic scale**
Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, Anna Goldenberg
Nature Methods (2014-01-26) https://doi.org/f5v9f5
DOI: 10.1038/nmeth.2810 · PMID: 24464287

38. **Recovering Gene Interactions from Single-Cell Data Using Data Diffusion**
David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, … Dana Pe'er
Cell (2018-07) https://doi.org/gdqdwj
DOI: 10.1016/j.cell.2018.05.061 · PMID: 29961576 · PMCID: PMC6771278

39. **Integrated analysis of multimodal single-cell data**
Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck III, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zagar, … Rahul Satija
Cold Spring Harbor Laboratory (2020-10-12) https://doi.org/ghhj3m
DOI: 10.1101/2020.10.12.335331