Semi-supervised identification of cell populations in single-cell ATAC-seq

Pawel F. Przytycki[1] and Katherine S. Pollard[1,2,3,4,5,6,*]

[1]Gladstone Institutes, San Francisco, CA, USA
[2]Chan-Zuckerberg Biohub, San Francisco, CA, USA
[3]Institute for Computational Health Sciences, University of California, San Francisco, CA, USA
[4]Institute for Human Genetics, University of California, San Francisco, CA, USA
[5]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA
[6]Quantitative Biology Institute, University of California, San Francisco, CA, USA
* Corresponding author: katherine.pollard@gladstone.ucsf.edu

Abstract

Identifying high-confidence cell-type specific open chromatin regions with coherent regulatory function from single-cell open chromatin data (scATAC-seq) is difficult due to the complexity of resolving cell types given the low coverage of reads per cell. In order to address this problem, we present Semi-Supervised Identification of Populations of cells in scATAC-seq data (SSIPs), a semi-supervised approach that integrates bulk and single-cell data through a generalizable network model featuring two types of nodes. Nodes of the first type represent cells from scATAC-seq with edges between them encoding information about cell similarity. A second set of nodes represents “supervising” datasets connected to cell nodes with edges that encode the similarity between that data and each cell. Via global calculations of network influence, this model allows us to quantify the influence of bulk data on scATAC-seq data and estimate the contributions of scATAC-seq cell populations to signals in bulk data. Using simulated data, we show that SSIPs successfully separates distinct cell types even when they differ in very few mapped scATAC-seq reads, with a significant improvement over unsupervised cell type identification. We apply SSIPs to scATAC-seq data from the developing human brain and show that supervising with just 25 differentially expressed genes from scRNA-seq enables the identification of two subtypes of interneurons not identifiable from scATAC-seq data alone. SSIPs opens the door to identifying high resolution cell types in single-cell open chromatin data, enabling the study of cell-type specific regulatory elements.

Keywords

Single-cell, chromatin accessibility, semi-supervised learning, graph diffusion, data integration

Introduction

Single-cell genomics is an exciting avenue to overcoming limitations of bulk tissue studies. Bulk data averages information over a heterogeneous collection of cells making it impossible to capture cell variability and cell-type specific transcriptomes and regulatory programs [1,2].

However, much work remains before we can accurately leverage single-cell assays to capture the diversity of cell types. In particular, these technologies struggle with low-resolution measurements featuring high rates of read dropout and few reads per cell [1,2]. Many methods have been developed to address these problems in single-cell expression data (scRNA-seq) [1,2]. However, methods that work reasonably well on scRNA-seq fail on scATAC-seq data. This happens because there are fewer reads per cell and the portion of the genome being sequenced is typically much larger than the transcriptome. Consequently, scATAC-seq has much lower coverage and worse signal-to-noise compared to scRNA-seq [3]. Several methods have been developed to increase the number of informative reads used per cell including CICERO [4], which aggregates reads from peaks that are co-accessible with gene promoters to emulate gene focused scRNA-seq data, and SnapATAC [5], which computes cell similarity based on genome-wide binning of reads. Other methods search for informative reads based on regulatory regions [6,7]. However, methods such as these are still often unable to detect known rare cell types in scATAC-seq data [3].

Our key insight is that cell types in scATAC-seq data need not be discovered *do novo* because concurrently generated and publicly available data contain information that can be leveraged. Our approach goes beyond co-clustering or jointly visualizing cells from scATAC-seq and scRNA-seq. Such methods attempt to detect cell types in scATAC-seq data by either mapping the data into the same projected space as scRNA-seq data [4,8] or by labeling cells in scATAC-seq to known cell-type expression profiles [9]. While these provide a promising avenue towards adding labels to clusters of cells observed in scATAC-seq data, they do not help to increase the resolution of cell type detection.

We propose that cell types derived from scRNA-seq data and reference cell atlases, combined with tissue and condition specific bulk measurements, can be integrated with scATAC-seq data in order to improve the resolution of cell types in that data using a semi-supervised learning method. In our approach, Semi-Supervised Identification of Populations of cells in scATAC-seq data (SSIPs), a subset of cells is sparsely labeled based on the accessibility of genes known to be expressed in specific cell types. We then propagate labels using a global graph diffusion algorithm through a network consisting of “supervising” nodes and cell nodes. Diffusion through cell-cell edges allows labeling information to indirectly influence cells with similar genome-wide open chromatin profiles even if that gene’s promoter is not detected as open. Our method only requires the tuning of a single parameter per external data set that determines the ratio of the weight of edges to that data relative to the weight of edges between cells. A major benefit of our method is that it computes the level of influence of each cell type from each external data source on every cell, thus allowing us to determine which data are the most informative.

Previous work has shown network-based methods to be a powerful tool for analyzing single cell sequencing data [10–12]. However, to the best of our knowledge, SSIPs is the first approach that flexibly integrates known cell types, cell types derived from matched scRNA-seq data, and tissue and condition specific bulk data to supervise the learning of cell types in scATAC-seq data. Using simulated data we demonstrate that our method is able to improve cell type resolution and then apply our method to scATAC-seq data from the developing human brain and show it successfully finds two subtypes of interneurons not previously identifiable from scATAC-seq data alone. The ability to increase cell resolution will allow for the identification of cell type specific regulatory elements.

Methods

Overview. We present SSIPs, a generalizable network model that improves the resolution of cell populations in scATAC-seq data by integrating it with other bulk and single-cell datasets (Figure 1A). Briefly, cells from scATAC-seq are nodes in the network, and edges between them encode information about cell similarity. A second set of nodes represents "supervising" datasets connected to cell nodes with edges that encode the similarity between that data and the cell. Information from the "supervising" data is propagated across all cells using a global graph diffusion algorithm.

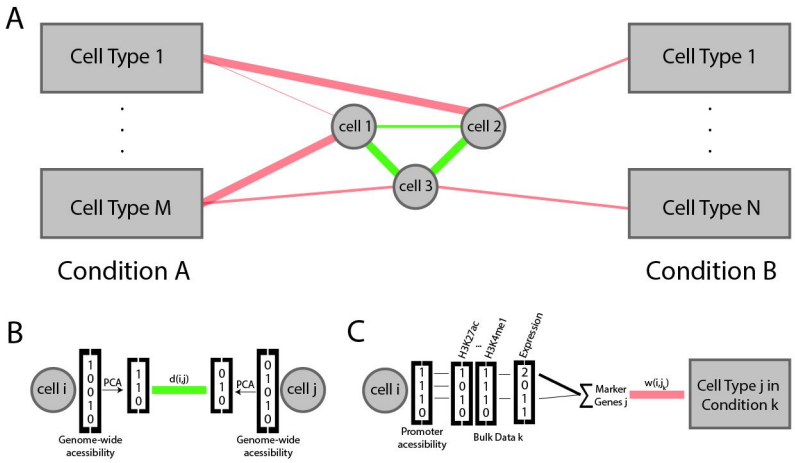

Figure 1. **Overview of Semi-Supervised Identification of Populations of cells in scATAC-seq data (SSIPs). A.** Cells (circles) are connected based on similarity of their scATAC-seq profiles (green edges). Their network is supervised by scRNA-seq and/or bulk data (rectangles) that pass information about cell types to scATAC-seq cells (red edges) using gene-based similarities. This general network structure provides a framework through which information from "supervising" data can be propagated to all cells. **B.** Edges between pairs of cells are weighted by the similarity of their binarized scATAC-seq profiles after dimension reduction. **C.** The edge weight connecting a cell to a cell type sums the expression of marker genes that have open promoters, optionally filtered through a Boolean combination of epigenetic marks.

Data encoding. We encode scATAC-seq data based on the format introduced by SnapATAC [5], a state-of-the-art unsupervised scATAC-seq analysis pipeline which produces counts of reads in each of p equal sized genomic bins (5kb resolution) for each of n cells. This p-by-n count matrix is binarized (open/closed). Cell types for each tissue and condition are represented by a binary g-by-c matrix of g marker genes for c cell types. RNA-seq is encoded as normalized expression per gene and epigenetic data is encoded as binary indicators for the presence of significant enrichment in each region.

Initial model. The p-by-n count matrix from a sample is used to compute the upper triangle of the n-by-n Jaccard similarity between all pairs of cells. The dimensionality of this matrix is reduced using a method such as principal components analysis (PCA) and then converted into network edges by computing the distance between all cells in the first m dimensions of the reduced space (m<n). The resulting unsupervised cell network is the low-resolution, initial model of cell clustering upon which we aim to improve using external data (Figure 1B).

Supervision of the cell network. The g-by-c matrices of marker genes and cell types for each condition or tissue are combined with bulk expression and epigenomic data in order to influence—or supervise—the similarity of cells in the initial network by adding edges that correspond to relationships between each external data source and a subset of cells (Figure 1C). Each edge from a cell type in a given condition is a weighted sum of the expression of marker genes in that cell type that have open promoters in the cell, optionally filtered through a Boolean combination of epigenetic marks depending on the availability of relevant bulk data. This weight is then normalized by the total accessibility of each cell. With this general approach, it is possible to add a large variety of external data to the model. Although these edges are sparsely connected to cells, the edges between cells distribute information. Supervision introduces one parameter, the label edge weight, per external data set which determines the ratio of the weight of edges to that data relative to the weight of edges between cells.

Network diffusion. To diffuse the information from all data sources across the network, we implemented a random walk with restarts. A unit amount of information is initialized at each node. Then at each time step, half restarts and the remainder propagates across each edge connected to the node, proportionally to edge weights. Even cells poorly annotated with external data will receive information about those annotations via cells that are similar. This algorithm is equivalent to an insulated heat diffusion graph kernel. To implement diffusion, we first compute a q-by-q walk matrix W encoding the fraction of information that must move to each neighboring node in each time step, where q is equal to the total number of nodes in the graph. This is 0 if the nodes have no edge between them and the fraction of total weight of edges for each node otherwise. In matrix notation the computation is $W=D^{-1}A$, where D is a diagonal matrix of the sums of edge weights for each node and A is the adjacency matrix representing the graph.

Influence matrix. Given this formulation of the walk matrix *W* and a non-zero restart probability α, the walk always converges to a stationary distribution. Due to this property, there is a closed form solution for the q-by-q influence matrix *F*, which defines the amount of information that reaches each node from each other node and is computed as $F = \alpha(I - (1-\alpha)W)^{-1}$. Prior work has examined how different settings of alpha distribute information to neighboring nodes and found that a restart probability between 0.4 and 0.6 encodes graph structure well with only minor variance in information in that range [13]. Based on this, we set our restart probability to 0.5. Since the network is sparse, performing the matrix inversions to compute *W* and *F* is feasible even for large numbers of cells and data types. Columns of *F* that correspond to "supervising" nodes represent the influence that information source has on each cell. We can also estimate the contribution of cells from the scATAC-seq data to each "superiving" piece of data by considering the influence each cell has on that data source, providing some indication of the composition of that bulk data.

Results

Generating simulated data. To investigate the potential benefits and limits of our model we tested it on simulated data. We generated artificial cells that emulate high quality scATAC-seq data by sampling from a pool of reads to build a binarized p-by-n count matrix of n cells and p bins that emulates those generated by the SnapATAC pipeline. First, for each cell we sample the number of total reads for that cell from the distribution of reads per cell we observed in real data (median 5,500 reads per cell). We then distribute those reads across bins proportionally to the distribution of reads per bin observed in real data. In order to introduce cell type similarity into the simulations data, we adjust the probabilities of reads falling into bins for each cell type by splitting bins evenly across cell types and adding a fixed percent of additional reads across those bins to each cell. We generated simulated matrices of 800 cells for five levels of cell similarity where each type was balanced with 400 cells each: no probability adjustment, and 1, 5, 10, and 100 percent additional cell type specific reads.

Quantifying performance. To measure performance we computed cell-to-cell homogeneity directly from the computed influence matrix *F* as the median ratio of information between cells within the same cell type to information between cells of different cell types. Calculation of cell homogeneity is only possible in simulations, where true cell types are known.

Supervision increases cell homogeneity. To test how much supervision increases cell homogeneity, we randomly selected 25% of cells in each of the two simulated cell types and added edges from those cells to two cell type nodes. By varying the weight of these label edges, we can tune the influence of supervising data and quantify the effect on performance (Figure 2A). When this weight is 0, there is no supervision, which is equivalent to running SnapATAC (purple area). As we increase the weight, SSIPs uses the supervising data more and passes

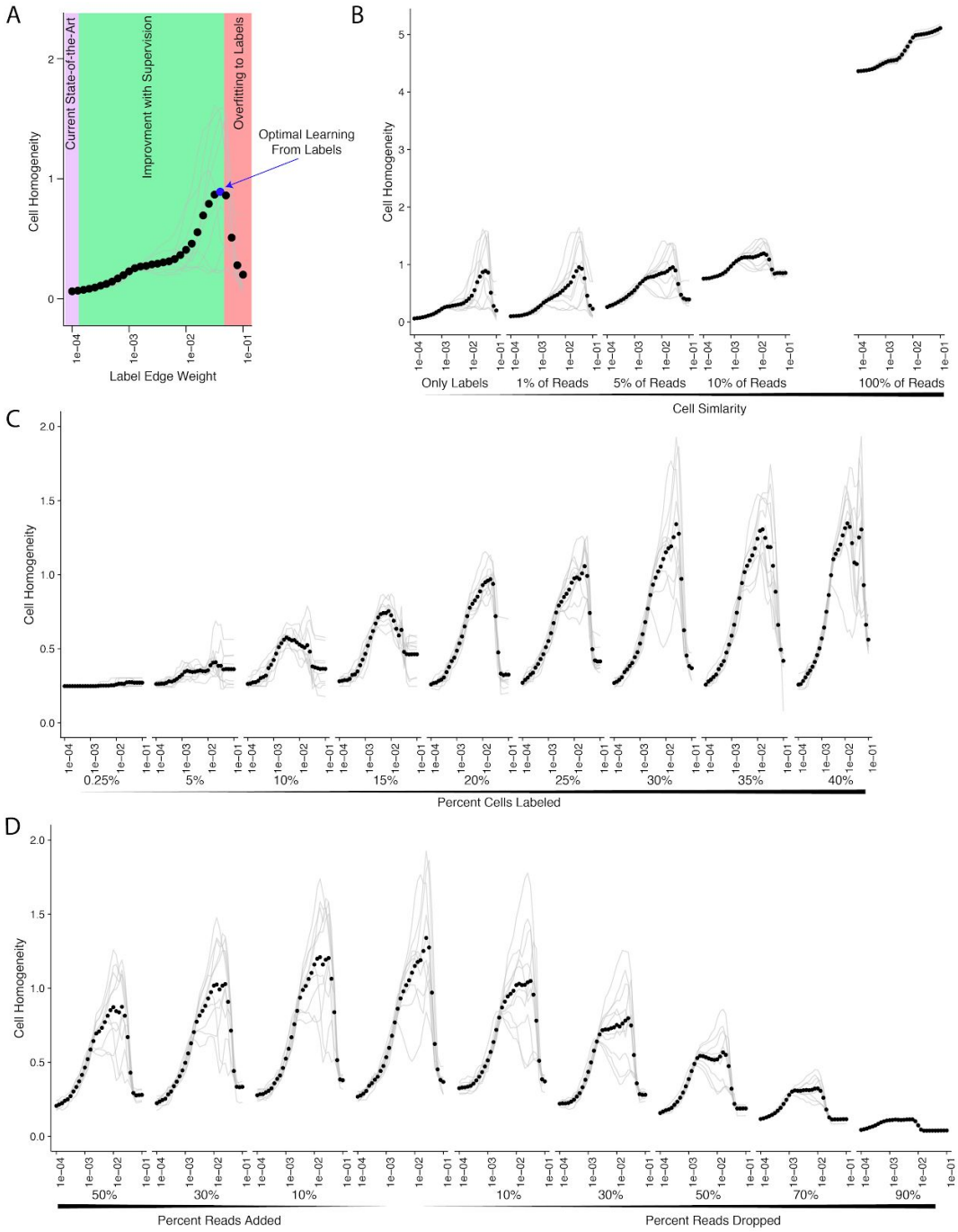

Figure 2. **SSIPs increases cell homogeneity in simulated data. A.** Increasing the weight of label edges (x-axis) increases the influence of supervising data. When this weight is 0, there is no supervision, which is equivalent to running SnapATAC (purple area). With increased supervision, cell homogeneity increases (green area) to a maximum (blue dot) and then declines (red area) as SSIPs begins to over-utilize the supervising data relative to cell-cell similarity leading to overfitting to that data. Gray lines represent each of 10 random selections of cells to add edges to with mean scores shown with black points. **B.** As expected, higher true cell similarity leads to higher initial cell homogeneity (first point on each curve). At every level of cell similarity, cell homogeneity increases with increasing label edge weight. The highest relative improvement is in cases with low cell similarity. **C.** As the percent of cells connected to "supervising" data increases, SSIPs has more power to improve cell homogeneity with 10% labeling serving as a minimum needed for significant improvement, and no additional benefits after 30% of cells are labeled. **D.** SSIPs is robust to randomly adding or removing reads from cells; even with as many as 50% of all reads dropped, peak cell homogeneity is still as high after supervision as without supervision when all reads are present.

information from connected cells to other similar cells in the scATAC-seq data. We expect that performance will increase (green area) and peak before maximal supervision (blue dot) and then decline (red area) as SSIPs begins to over-utilize the supervising data relative to cell-cell similarity leading to overfitting to that data. This is indeed what we observe. In order to generate robust estimates of cell homogeneity, we repeat the random selection of cells to add edges to 10 times (gray lines) and compute the mean cell homogeneity at each label edge weight (black points). We ran this same set of tests across the five simulated matrices representing varying levels of true cell similarity (Figure 2B). As expected, as true cell similarity increases, cell homogeneity increases even without any supervision (first point on each curve). At every level of cell similarity, cell homogeneity increases with increasing label edge weight. In the final case, where cells from different cell types are very distinct from each other, initial cell homogeneity is already so high that only relatively minor improvements to cell homogeneity are possible. Overall we observe the highest relative increase in cell homogeneity when the cells types are hard to distinguish using unsupervised methods.

Increased labeling improves cell homogeneity. The ability to supervise the learning of cell types in scATAC-seq data is highly dependant on the number of cells that can be initially labeled using edges to “supervising” data. In order to establish bounds on how many cells must be labeled per cell type to successfully increase cell homogeneity we measured improvements to cell homogeneity when labeling between 0.25% (1 cell) and 40% of cells when cell types initially differ in 5% of reads (Figure 2C). We observe that as few as 10% of cells being labeled is sufficient for improvement in cell homogeneity and that there is no further improvement after 30% of cells are labeled. To determine if this is a function of the number of cells labeled or the percent of cells, we ran the same test with a simulation of half as many cells and observed similar limits with 10% labeling sufficient for improvement and no further benefits past 35% of cells being labeled. These bounds establish the limits on how many cells must be labeled in each cell type to effectively identify those cell types in scATAC-seq data.

Label propagation is robust to noise. Given that scATAC-seq is very noisy, it is important for cell populations to be identifiable even if a large portion of reads are missing or if large regions of accessibility were detected. Using simulations, we estimated how sensitive our method is to noise in cell accessibility profiles by randomly adding or removing reads from cells while maintaining 30% of cells as labeled (Figure 2D). Even with as many as 50% of all reads dropped, peak cell homogeneity is still as high after supervision as without supervision when all reads are present. In the opposite direction, when random noise is added we only observe slightly diminished cell homogeneity indicating that our method is robust to cells with broad accessibility.

Semi-supervised learning identifies rare interneuron subtypes. We leveraged mid-gestation, human telencephalon scATAC-seq data from psychENCODE [14] to test the

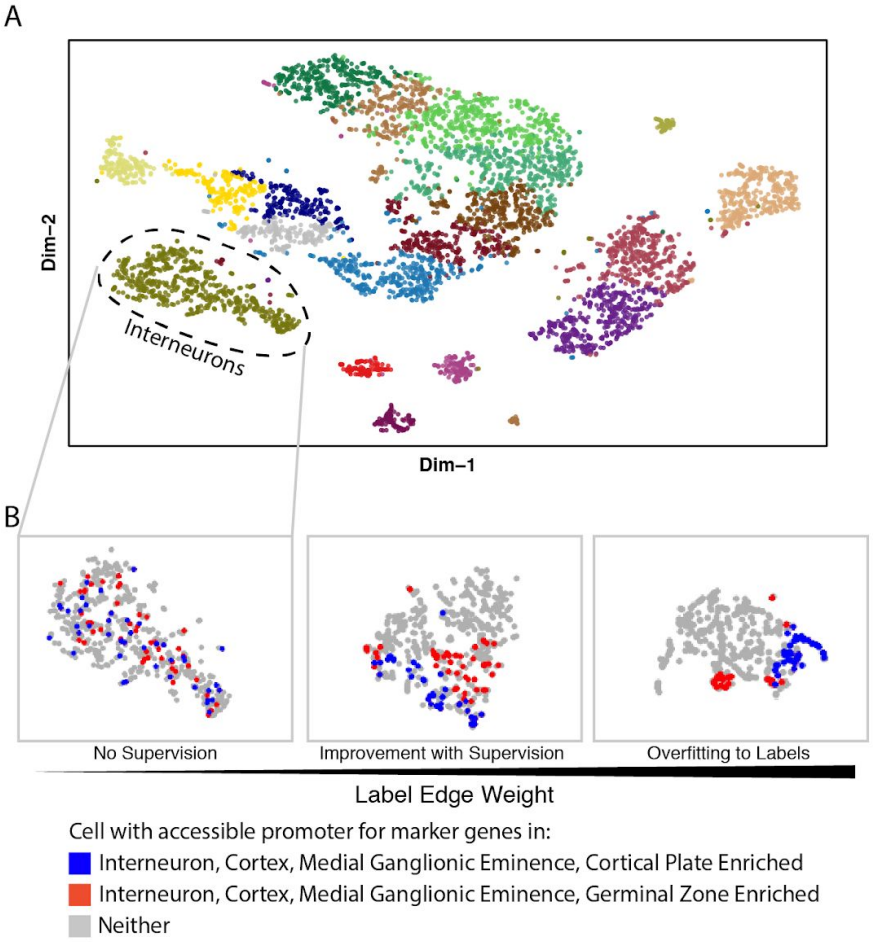

Figure 3. **Semi-supervised learning identifies rare interneuron subtypes. A.** Overall, MGE interneurons cluster together based on scATAC-seq profiles**. B.** However, based on promoter accessibility for marker genes, interneuron subtypes are not resolved within this cluster (blue versus red, left panel). When we introduce supervision using SSIPs, the two subtypes become clearly distinct components of the initial interneuron cluster (middle panel). Finally, when the weight on the label edges becomes too high, we observe overfitting in which labeled cells begin to cluster only with other labeled cells (right panel).

ability of our model to resolve cell types in real data. Previous work generated a cell type atlas in a similar context based on extensive analysis of scRNA-seq data [15]. This atlas observed two types of interneurons in the medial ganglionic eminence (MGE): Type 1 (cortical plate enriched) and Type 2 (germinal zone enriched). However, MGE interneurons formed a single cluster in scATAC-seq when run through the standard SnapATAC pipeline (Figure 3A). We then selected 25 marker genes that are at least 2-fold overexpressed in an interneuron subtype versus all other cell types and found that cells with accessible promoters for these marker genes are distributed throughout the interneuron cluster (Figure 3B, left). We note that about 15% of cells are labeled in each cell type if we assume that the interneuron cluster is composed

approximately equally of the two types, which is sufficient to identify cells types according to our simulated results. Next we computed the influence matrix *F* across many levels of label edge weight, selected the first two columns of *F* (these represent the influence of the two cell types nodes on all cells), and added those as additional columns of the dimensionality reduction produced by SnapATAC before re-projecting the data using default parameters. We find that at good levels of label edge weight, labeled cells begin to separate in projected space (Figure 3B, middle) but that once the label edge weight is too high we observe overfitting in which labeled cells begin to cluster only with other labeled cells (Figure 3B, right). The resulting successful separation of cells with markers for different interneuron subtypes shows that the additional information from scRNA-seq is sufficient to separate interneurons from scATAC-seq into the two known subtypes. This finding establishes the feasibility of semi-supervised, network models for resolving cell subpopulations in scATAC-seq data.

Discussion

Using SSIPs for semi-supervised identification of cell populations we have shown that external data about known cell types can help overcome the low signal-to-noise ratio present in scATAC-seq data. This strategy is extremely powerful as there are already vast amounts of bulk sequencing, scRNA-seq, and compiled cell atlas data for tissues, organoids, and cell lines related to samples where scATAC-seq is being performed. Large scale integrations of such data are the primary application of our model.

An exciting extension of SSIPs will be to integrate multiple scATAC-seq samples into a single graph. This can be done by computing edge weights between cells from different samples and allowing all cells to connect to external data nodes. This will allow us to measure cell similarity across samples. A compelling benefit of multi-sample integration is that the edges presented by external data may also help with batch correction by pulling similar cell types together.

An important caveat to our semi-supervised learning model is that cell types with no known labels will begin to look more similar to each other as label edge weight is increased. For example, if a group of cells that are initially similar to each other is composed of many cell types, but we only have known cell labels for two of those types, those two will be pulled away from the remainder of the cells. The middle panel of Figure 3B, where a new cluster of completely unlabeled cells appears to be forming, may be an example of this behavior. Easy-to-collect bulk expression and epigenetics data for more cell types and tissues, as well as the growing number of scRNA-seq data sets, should mitigate this problem.

Once cell types have been identified in scATAC-seq data, usually the next downstream step is to call cell-type specific regulatory elements. This is generally done by re-agregatting reads across cells that are assigned the same cell type and then calling peaks [4,5,7]. For this aggregation to work well, given the level of noise in scATAC-seq data, correct assignment of cell populations is very important. SSIPs can therefore serve as a powerful step within the traditional pipeline of scATAC-seq analysis, where it can simply be inserted as an extra step after the initial dimensionality reduction is completed and before further downstream analysis.

References

[1] Hwang B, Lee JH, et al. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp. Mol. Med. 2018;50:96.

[2] Haque A, Engel J, et al. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017;9:75.

[3] Chen H, Lareau C, et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data [Internet]. Bioinformatics; 2019 [cited 2019 Oct 31]. Available from: http://biorxiv.org/lookup/doi/10.1101/739011.

[4] Pliner HA, Packer JS, et al. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol. Cell. 2018;71:858-871.e8.

[5] Fang R, Preissl S, et al. Fast and Accurate Clustering of Single Cell Epigenomes Reveals *Cis* -Regulatory Elements in Rare Cell Types [Internet]. Bioinformatics; 2019 [cited 2019 Oct 31]. Available from: http://biorxiv.org/lookup/doi/10.1101/615179.

[6] Bravo González-Blas C, Minnoye L, et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. Nat. Methods. 2019;16:397–400.

[7] Schep AN, Wu B, et al. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods. 2017;14:975–978.

[8] Stuart T, Butler A, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177:1888-1902.e21.

[9] Pliner HA, Shendure J, et al. Supervised classification enables rapid annotation of cell atlases. Nat. Methods. 2019;16:983–986.

[10] Johansen N, Quon G. scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. Genome Biol. 2019;20:166.

[11] Tarashansky AJ, Xue Y, et al. Self-assembling manifolds in single-cell RNA sequencing data. eLife. 2019;8.

[12] Wang Y, Hoinka J, et al. Subpopulation Detection and Their Comparative Analysis across Single-Cell Experiments with scPopCorn. Cell Syst. 2019;8:506-513.e5.

[13] Leiserson MDM, Vandin F, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat. Genet. 2015;47:106–114.

[14] Wang D, Liu S, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018;362:eaat8464.

[15] Nowakowski TJ, Bhaduri A, et al. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science. 2017;358:1318–1323.