Computational Tools for Characterizing Cellular Profiles using Multiple Measures of Cellular Identity

Specific Aims

The capabilities of single-cell sequencing modalities are expanding at an exponential rate. Novel spatial transcriptomics protocols[1–4] provide transcriptomic profiles for cells while preserving information about the cell's origin within a tissue. New dual-omics technologies allow the simultaneous capture of epigenomic and genomic expression profiles from the same cell[5,6]. However, despite the accelerated progress of single-cell sequencing techniques, no current modality is capable of capturing epigenomic, genomic, and spatial information from the same cell. The transcriptome and epigenome are instrumental in shaping the development and maintenance of a cell's morphology and function. However, a cell is also profoundly affected by where it resides within a tissue, as this affects its exposure to extracellular products and signals. Consequently, without considering the epigenomic, genomic, and spatial context of a cell, our understanding of cellular identity remains disjointed. Without a clear definition of a cell type, researchers struggle to achieve insight into how various cell types are developed and maintained, as well as how these processes might break down in the context of disease. Motivated to establish cellular profiles defined by these three facets of cellular identity, we propose a workflow that jointly interrogates the intimate relationship of these measures of cellular identity throughout development.

Aim 1: Develop a novel dataset integration method that includes spatial coordinate information for improved integration results.

Our current method, LIGER, utilizes integrative nonnegative matrix factorization (iNMF) to achieve dataset integration[7]. I propose updating the current iNMF algorithm to incorporate graph regularization[8] , leveraging both spatial and molecular data when defining cell types. Graph Regularization uses the local invariance principle, the idea that two points near each other are more likely to be similar. For dataset integration, this would leverage the knowledge that two cells anatomically near each other would be more likely to be similar, as they are more likely to share a tissue of origin and similar extracellular signals. We hypothesize that the incorporation of this additional information will lead to improved integration results, and better characterization of cell type distinctions for cells with spatial coordinates. The benefits of using LIGER to integrate the spatial transcriptomic dataset, STARmap[1] , with a scRNA-seq[9] dataset has already been demonstrated[7]. By using Graph Regularized iNMF to integrate these same datasets, we can quantify the increase in algorithm performance using purity and Adjusted Rand Index measures. The increase in the quality of dataset integration would directly reflect the benefit of including the spatial information into the matrix decomposition.

Aim 2: We will impute scRNA profiles and predict scATAC profiles within a spatial context.

Previously, K-nearest neighbor (KNN) algorithms have been used to successfully impute transcriptomic values for genes not measured by a spatial transcriptomic modality[7,10,11]. Dual-omics datasets, characterized by epigenomic and transcriptomic measures from the same cell, provide linked cell profiles that are the key to inferring epigenetic profiles for cells with spatial coordinates. By leveraging a dual-omics dataset, we propose that a modified KNN approach can also be used to predict epigenomic profiles for the cells with spatial coordinates, providing some of the first spatial context for epigenomic measures. To validate our method, we will use two peripheral blood mononuclear cell dual-omics datasets[12,13]. First, we will first integrate the dual-omics datasets using their scRNA-seq profiles. Then we will use KNN to predict the scATAC profiles for dataset one using the scATAC profiles from dataset two. The success of our prediction can be measured using the withheld scATAC profiles as ground truth. Aim 2 is independent of Aim 1, as Aim 2 can be applied to current LIGER integrations.

Aim 3: Identify pathways that are differentially regulated spatially and temporally within the mouse prefrontal cortex.

We hypothesize that using transcriptomic and epigenetic profiles with spatial resolution will identify spatially differentiated pathways of expression in the prefrontal cortex, a region that undergoes rapid development in the weeks prior to birth[14]. Proper development of this region is key, as inappropriate formation is associated with impaired cognitive ability and psychiatric disorders[15]. Using Spatially-resolved Transcript Amplicon Readout mapping, STARmap[1], as well as the 10X dual-omics kit, we will obtain both spatial transcriptomic and dual-omics expression profiles of the mouse frontal cortex at three key stages of embryonic development. We will use our novel iNMF algorithm to integrate the datasets, and our novel KNN approach to predict epigenetic expression profiles for the STARmap cells. We will use SPARK[16] to identify spatially expressed genes and differential chromatin accessibility, and can also use TimeReg[17] to examine transcriptome and epigenome differences between development stages.

Background and motivation

Each cell of the human body has the same underlying genetic code, or DNA, but the way this DNA is read changes dramatically from cell to cell. These different interpretations of DNA result in the incredible

diversity we observe in the morphology and function of cells across the human body. While such complexity is necessary to support life, researchers struggle to understand the minute factors that characterize each cell type. To better distinguish each cell type, researchers examine individual attributes that contribute to a cell's identity: the epigenome, the transcriptome, as well as a cell's location within a tissue. Each of these measures provides a different lens through which researchers can examine similarities and differences between cells, and consequently construct cellular profiles. These profiles describe patterns of expression that are typically presented by a certain cell type, and can likewise offer information about how these patterns might vary as a result of a cell's spatial position.

The ability to examine and define cell types at high resolutions has been largely driven by the multitude of advances in single-cell sequencing technologies. scRNA-seq allows cell types to be defined based on their transcriptomic profiles, and scATAC-seq has been similarly instrumental in understanding how epigenetic status contributes to cell type distinctions. Breakthroughs in spatial transcriptomic technologies offer unique insights into how transcriptomic profiles vary by location, and novel dual-omics protocols have been developed to capture epigenomic and transcriptomic from the same barcoded cell.

Despite the rapid progress of single-cell technologies, no technology is currently capable of simultaneously capturing epigenomic, transcriptomic, and spatial facets of cellular identity. This leaves current cell type characterizations fractured and disjointed, as researchers struggle to understand the intricate dynamics between these different measures of cell identity. Without a thorough understanding of the intimate relationships between the epigenome, transcriptome, and spatial location, global characterization of cell types is intractable, and it is difficult to determine what perturbations of cellular expression are expected, and what levels have the potential to result in pathogenic outcomes. Consequently, a key next step in single cell data analysis is the development of computational methods motivated to establish a better understanding of how these three facets of cellular identity interact.

The innovative tools and novel applications we aim to develop are extendable to a growing multitude of datasets; however, in Aim Three, we focus on the developing prefrontal cortex, a key control of higher cognitive ability. We prioritize investigating the development of the prefrontal cortex because perturbations of its development are responsible for the cognitive deficits of neurodevelopment disorders such as intellectual disability, Attention Deficit Hyperactivity Disorder (ADHD), schizophrenia, and autism spectrum disorders[18]. These diseases present a remarkably complex etiology, which parallels the equally intricate developmental events that shape the prefrontal cortex. The medications and treatments offered to treat such disorders are not curative, strongly motivating the necessity of preventative action. Researchers have suggested that intervention into unbalanced development programs might rescue normal prefrontal cortex development[18]. However, before such interventions are possible, it is first necessary to understand the delicate process that orchestrates cortex development. Therefore, understanding how the epigenome, transcriptome, and spatial location all contribute to the mechanisms of cortex development is a crucial next step for untangling the etiology behind neurodevelopmental disorders. Motivated to address this knowledge gap, we focus on untangling the mechanisms underlying the development of the prefrontal cortex.

Innovation: Despite recent advances in spatial transcriptomic technologies, most are not yet capable of providing the same transcriptomic resolution as scRNA-seq experiments[19]. Consequently, researchers choosing between spatial transcriptomic methods often must weigh trade-offs between the optimal number of genes, the desired number of cells, and the necessary resolution most appropriate for their question of interest[20]. The limitations of these assays inhibit the ability of researchers to characterize cell types within a spatial context[19]. Therefore, we propose a novel dataset integration approach for single cell spatial transcriptomic and scRNA-seq datasets that will allow for a refined characterization of cells with spatial coordinates.

Current spatial technologies and bioinformatics tools focus largely on the transcriptome[1,4,21–23], with the only spatial epigenetic assay lacking single cell coordinate information[24]. In order to explore how the relationship between the epigenome and the transcriptome are perturbed by location within a tissue, we propose a method that allows for the prediction of chromatin accessibility profiles for cells with spatial coordinates.

Aim 1: Develop a novel dataset integration method that includes spatial coordinate information for improved integration results

Rationale: Characterizing cell types is already a major obstacle in analyzing single cell sequencing data, but it is particularly challenging for spatial transcriptomic datasets, as transcriptomic profiles for cells captured by spatial transcriptomic protocols typically have a lower resolution than that obtained by scRNA-seq[19]. To address this difficulty, computational methods have primarily taken two distinct approaches. Methods such as

Seurat[11] and LIGER[7] integrate spatial transcriptomics data with scRNA-seq data, using the higher resolution scRNA-seq data to establish more refined profiles then would be possible using the spatial transcriptomic data alone. Although this method has been shown to increase the resolution of cell type assignments[7,25], it ignores any information relevant to spatial location when performing dataset integrations. Methods such as SPICEMIX[22] and FICT[23] exemplify the second approach to cell type characterization, jointly using gene expression and spatial information from a single spatial transcriptomics experiment to resolve cell types. Such methods have shown improved performance over methods that characterize the cells by gene expression profiles alone[22,23], yet these methods fail to leverage any external information from higher resolution datasets. Because both strategies individually improve cell type resolution, we hypothesize that using them in conjunction will additively increase our ability to define cell types for spatial transcriptomics data.

Methods: Our current approach, LIGER[7], uses integrative non-negative matrix factorization (iNMF) to integrate datasets across experiments and modalities. On numerous occasions[7,25], LIGER has successfully improved resolution of a spatial transcriptomic dataset through integration with a scRNA-seq dataset[7,25]. The current iNMF equation decomposes each dataset without accounting for any spatial information. We posit that including spatial coordinate information during dataset integration will result in improved integration results, and consequently better resolved cell type characterizations. To incorporate the spatial information when performing dataset integrations, we propose introducing Graph Regularization into the iNMF algorithm **(Fig 1)**.

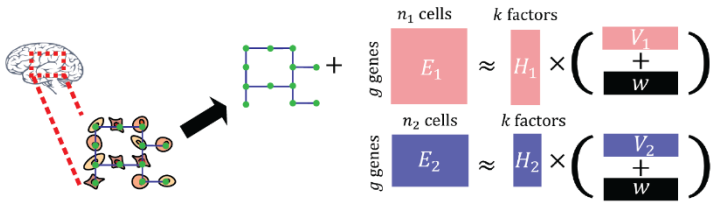

Figure 1: The GRiNMF algorithm will use the spatial coordinates, encoded as a graph, in each iteration of the matrix decomposition. This produces metagenes that are informed by the cell's placement within a tissue, and improves the quality of dataset integration

Graph Regularized Non-Negative Matrix Factorization[8](GNMF) is theoretically driven by the invariance principle. The invariance principle captures the idea that two objects close to each other in space are more likely to be similar, and its implementation in GNMF forces the matrix decomposition to respect relationships defined in a graph structure. To include spatial information into the matrix decomposition, GNMF defines a matrix, M, where M is n by n, where n is the number of cells. The entries of M are populated with a weighting scheme that is problem dependent. We initially chose the 0-1 weighting scheme, where for cells i, j:

$$M_{i,j} = \begin{cases} 0 \; if \; i,j \; are \; not \; K \; nearest \; neighbors \\ 1 \; if \; i,j \; are \; not \; K \; nearest \; neighbors \end{cases}$$

This weighting scheme is simple to calculate, and limits the number of free parameters for the optimization to two (K, λ). D is a diagonal matrix populated by summing the rows of M, so we may define the graph laplacian as $L = D - M$. Incorporating this weighted matrix into the penalty term of the original iNMF equation, we derive:

$$\arg\min_{H_i,V_i,W,\,L\geq 0} \sum_i \|E_i - H_i(W+V_i)\|_F^2 + \lambda_i Tr\big((W+V_i)^T L(W+V_i)\big) + \lambda \sum_i \|H_i V_i\|_F^2 \quad (1)$$

The novel iNMF algorithm, GRiNMF (1), still decomposes the individual datasets into shared metagenes (W), dataset-specific metagenes (V_i), and cell factor loadings (H_i), but also accommodates the spatial information (L) with each iterative update. To ensure that the spatial penalty is constrained to only operate on datasets with spatial coordinates, we vectorize λ_1, s.t. $\lambda_{1,i} = 0$ for datasets without spatial coordinates. The λ_2 penalty term is not vectorized, as it accounts for the overestimation of $W$and the underestimation of V_i that occurs as a result of homogeneity between the gene expression values of the datasets. Similar to the optimization problem posed by iNMF, we can solve the GRiNMF optimization problem with block coordinate descent[26].

Datasets: To validate our method, we will use the spatial transcriptomics dataset of the mouse frontal cortex captured using the STARmap[1] technology (2,522 cells; 1,020 genes), as well as the scRNA-seq dataset of the mouse frontal cortex captured using DROPviz[9] (28,366 genes; 71,639 cells).

Experimental Design: We will normalize and scale the datasets, and select genes common to both datasets. We will select the highest performing λ for each algorithm by observing performance over for a range of λ. Using similar heuristic methods, we will also optimize our choice for K as that which yields the highest average ARI and Purity scores. To quantify the advantage of GRiNMF over iNMF, we will integrate the STARmap and DROPviz datasets with ten different random initializations and calculate the purity and adjusted Rand Index (ARI) scores across a span of Louvain resolutions. For such calculations, we will use the DROPviz labels as ground truth, as these cell labels are considered high-quality annotations.

Expected Outcomes: We expect there to be a significant increase in both the ARI and Purity scores when using the GRiNMF algorithm, as compared to standard iNMF. Such results would confirm that including the spatial information in the dataset integration resulted in more accurate cluster assignment and consequently more refined characterization of cells with spatial coordinates. GRiNMF is a tool with widespread applicability independent of Aims Two and Three, as it can be used with any spatial transcriptomic and single cell dataset integration. Additional analysis that could benefit from GRiNMF include the integration of DROPviz data with existing spatial transcriptomic datasets gathered using the osmFISH[2], MERFISH[21], and Slide-seq[4] protocols.

Potential Issues and Alternative Approaches: One challenge that we anticipate facing is choosing appropriate values for K and λ.While heuristically optimizing algorithm performance is an initially time-consuming approach, we anticipate that after examining several datasets from diverse spatial transcriptomic technologies, it would be possible to develop intuition for an appropriate default parameter selection. If GRiNMF fails to improve dataset integration, an immediate strategy would be to explore alternative weighting options for the W matrix. Specifically, heat-kernel weighting adjusts the contribution of each nearest neighbor using spatial proximity, with the closest neighbors having correspondingly heavier weights. We opt to begin with the 0-1 weighting scheme because the heat-kernel weighting scheme entails defining an additional free parameter, the scaling factor σ, which has been shown to largely affect algorithm performance[8]. In the event that GRiNMF fails to outperform iNMF despite exploring all avenues of parameter adjustments and weighting schemes, the remaining aims may still be addressed using the original iNMF approach.

Aim 2: We will impute scRNA profiles and predict scATAC profiles within a spatial context.

Rationale: The K-nearest neighbor (KNN) algorithm is a heavily utilized machine learning algorithm that assumes that the close proximity of objects within space will imply an elevated measure of similarity between the objects. Within the context of single-cell dataset analysis, KNN algorithms leverage the idea of spatial proximity to impute missing gene expression values. Both Seurat[11,27] and LIGER[7], two popular dataset integration methods, use a type of KNN algorithm on integrated datasets to infer expression values for genes that are not present in all datasets. This method has particular utility when integrating high quality data measured across many genes (i.e. scRNA-seq) with data defined by fewer measured genes, a characteristic typical of spatial transcriptomics data. The use of KNN in such a paradigm provides a means of imputing the expression values of genes originally unmeasured within space (**Fig. 2**).

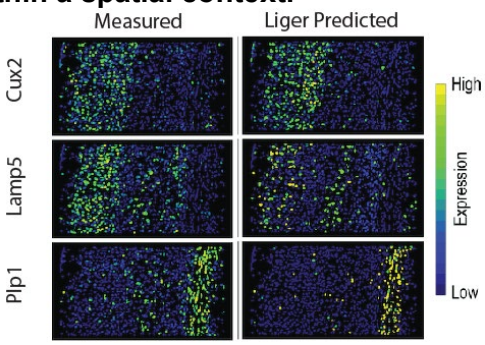

Figure 2: Comparison of the withheld ground truth expression levels (left) from a spatial transcriptomic experiment with the imputed gene expression values from LIGER's KNN algorithm (right) demonstrates the utility of KNN to impute expression profiles within a spatial context.

Although KNN imputation can be remarkably effective, it has limited utility in cross-modality prediction because the methods used to align these datasets assume that similarity between the distinct measures will reflect a common cell type. Yet, epigenomic changes frequently foreshadow gene expression, especially in regions undergoing rapid development[28,29]. The resulting decoupling between the expression values of different measures of a cell's identity can affect clustering. For instance, when integrating the scATAC and scRNA expression levels from the same cell, the lack of succinct correlation between these differing measures of cell identity results in imperfect pairing of scRNA and scATAC expression states (**Fig. 3**). The inconsistency in correctly matching scRNA and scATAC profiles that originate from the same cell illustrates a key barrier in using KNN to predict cross-modality expression profiles.

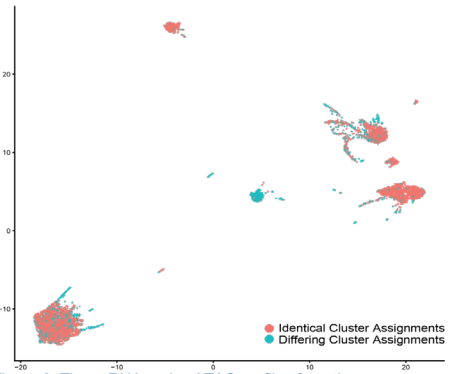

Figure 3: The scRNA and scATAC profiles from the same barcoded cells have been integrated as if they came from distinct experiments. The epigenetic and transcriptomic profiles that correspond to the same cell, but are still placed in different clusters, are colored in blue.

The advent of multi-omic sequencing, which jointly captures scATAC-seq and scRNA-seq from the same barcoded cell, provides a unique opportunity to explore KNN prediction across measures of cellular identity. By integrating the transcriptomic profiles of two datasets, it is possible to use the linked epigenetic profiles from the dual-omics data to infer scATAC expression profiles for the dataset defined only by transcriptomic measures. Consequently, the nearest neighbors are identified using a consistent measure of cellular identity for both datasets, ensuring that the cells identified as nearest neighbors share a similar molecular state. This has profound potential for spatial transcriptomic datasets, as the only protocol for providing scATAC profiles with spatial resolution lacks single cell coordinates[24].

The paired use of KNN and multi-omic datasets to predict scATAC profiles for transcriptomic data has previously been addressed only as baseline measurement of performance, where a basic KNN algorithm was still highly competitive with deep learning methods[30]. Motivated by the quality predictions of the unoptimized approach, we propose a refined KNN approach: Weighted Shared Nearest Neighbor (WSNN). Previous work has found that weighting the contribution of each nearest neighbor by the ratio of shared nearest neighbors improved algorithm performance[31,32], and we adapt this approach for scATAC-seq data. To upweight the contribution of more similar cells, we adjust the contribution of each neighbor on the number of shared nearest neighbors. By identifying the K-nearest neighbors of each cell using the factor loadings from iNMF, we reduce the noise introduced by irrelevant genes. We expect these innovations will elevate the performance of WSNN to a level that supersedes the established deep learning model[30].

Datasets: To quantify algorithm performance, we will use two multiome PBMC datasets from 10x Genomics. Dataset one profiles 10,000 cells[12], while dataset two provides 3,000 cells[13].

Experimental Design: Using only the transcriptomic information, we will integrate the PBMC data sets using LIGER[7]. For this validation experiment, we will not utilize GRiNMF, as we are performing our algorithmic validation on non-spatial data. Indeed, this highlights that WSNN's prediction capabilities will be capable of being utilized on data without spatial coordinates as well. When integrating the two PBMC datasets, we will follow the established LIGER pipeline, which includes data normalization, scaling, matrix factorization, and quantile normalization. Consequently, we will have normalized factor (H) loadings for each cell. To perform WSNN, we will utilize P_j, the chromatin accessibility profiles of dataset two. We can define P_j as either a binary or count matrix. When we define the chromatin accessibility count matrix using gene-level counts[7], we sum the number of reads overlapping the gene body or promoter region of a gene to yield a count matrix for P_j. Alternatively, we can define P_j using binary values to denote the presence or absence of a peak within a single cell. We will explore the utility of WSSN in making predictions using both definitions:

For cell i in dataset one (3K cells), we will identify it's K-nearest neighbors (K_i) from dataset two (10K cells). For cell $j \epsilon K_i$, we identify the number of shared nearest neighbors between i, j, such that $S_{i,j} = |K_i \cap K_j|$. The predicted profile, P, for cell i is then:

$$P_i = \frac{\sum_{j=1}^{K} S_{i,j} * P_j}{\sum_{j=1}^{K} S_{i,j}}$$

When using a binary expression of chromatin accessibility, each entry of the vector of P_i is the weighted percentage of shared nearest neighbors who also express this peak. Therefore, for binary accessibility values we set a global threshold, T, such that, for peak z in cell i,

$$Z_i = 1 \; if \; P_i \geq T.$$

Otherwise, $Z_i = 0$

The contribution of each nearest neighbor to the final prediction value is weighted by the similarity of the neighborhoods between the cell being predicted and cells being used for prediction. Other methods utilizing a thresholding approach on scATAC data found 0.1 to be an optimal value for T, although we will explore a range of values. We will use the Area Under the Receiving Operating Characteristic (AUROC) to quantify algorithm success when predicting binary count matrices[30], and the Pearson and Spearman correlation to examine the success of prediction for gene-level count predictions. Additionally, we will individually perform matrix factorization on the true scATAC expression values and the predicted ATAC values for each cell. Using the clustering results of the true scATAC expression values as ground truth labels, we can evaluate the purity and ARI of the predicted expression values. As K is a nontrivial free parameter of WSNN, we will examine its effect in algorithm performance over a range of K, s.t. $K \epsilon \{1,5,10,15,20,25\}$, as these are values standard for KNN exploration and imputation[7,11,30].

The alternative existing algorithm for scATAC prediction is BABEL[30], a recently proposed deep learning method for translating between scRNA-seq and scATAC-seq expression values at single-cell resolution. BABEL has shown decreased performance when predicting mouse profiles, a primary model organism for gathering spatial transcriptomic data. Additionally, BABEL performance on input with a reduced number of genes, an attribute of most spatial transcriptomic datasets, has not been quantified, and BABEL only works to predict binary accessibility measures. To evaluate BABEL performance on datasets with reduced genes, we will train and evaluate BABEL capabilities of predicting scATAC profiles for the 3K PBMC cells using both the full gene set, as well as the 1000 most variable genes. This will quantify how BABEL's prediction accuracy is affected by a reduced number of genes. We will likewise calculate the AUROC, Purity, and ARI scores for BABEL predictions.

Expected Outcomes: We expect WSNN to yield predicted scATAC values that have a higher AUROC, purity, and ARI scores than that of the existing method, BABEL. This would indicate that WSNN has substantial potential to accurately predict scATAC profiles from transcriptomic data, and that it would be the most appropriate method for predicting epigenomic profiles for spatial transcriptomic data. Since BABEL is incapable of predicting non-binary count values for accessibility data, we will not be able to compare correlation coefficients when predicting count matrices using WSNN.

Potential Issues and Alternative Approaches: Although previous KNN approaches have been applied to scATAC-seq data[30], the increased sparsity of the scATAC-seq data is a concern. If the success of prediction using WSSN is minimal, signal strength enhancement using scOPEN[33] can be explored. With the rapidly increasing number of both multiomic and spatial transcriptomic datasets, WSNN will have increasing utility. In the event that Aim Three is not completed, WSNN can still be used to generate scATAC-seq profiles with spatial coordinates by leveraging integrations of the SNARE-seq[6] multi-omic dataset with either the osmFISH[2] or STARmap[1] datasets. In such scenarios, predicted scATAC-seq profiles might be aggregated into pseudo-bulk measures and compared with existing bulk ATAC-seq assays[34] of the mouse frontal cortex to confirm the feasibility of the predicted values. We expect WSNN to outperform BABEL when predicting scATAC profiles from transcriptomic data. However, in the event that WSNN underperforms BABEL, we will use BABEL to predict scATAC profiles for the cells with spatial coordinates.

Aim 3: Identify differentially pathways that are differentially regulated spatially and temporally in the mouse prefrontal cortex.

Rationale: Previous work focused on disentangling development programs of the prefrontal cortex have primarily concentrated on the transcriptional dynamics of the developing brain[35], however, recent investigations have illuminated key roles of various chromatin mechanisms in neurogenesis[36]. Epigenomic regulations are hypothesized to help control migration of the neurons to their final positions, as well as aid in the establishment of synaptic connections with other neurons. With such critical roles, it is hardly surprising that variations in chromatin accessibility and other gene regulation mechanisms early in development is suggested as a key driver of developmental differences observed in adulthood[36]. Both the transcriptomic and epigenomic profiles of cells thus provide critical insights into cell fate decisions and development, and it is necessary to assess how each of these aspects of cellular identity jointly collude to produce the observed trajectory of both cell and brain development. As an individual cell's expression of these attributes will also be permuted by its surrounding environment and position in the developing cortex, examining cells using multi-omic measures and spatial coordinates provides the means to examine gradient fluctuations in gene expression, and further refine cell type characterizations. Specifically, it has been previously demonstrated that some sub-populations of cells are distinguishable solely by differences observed in a single modality. By examining both epigenomic and transcriptomic profiles within space, we can distinguish differences between cells that may be indistinguishable using only a single measure of cell identity, resolving more minute characterizations of cell types. Examining how these profiles change over the course of development establishes intuition as to critical gene regulatory modules and transcription factors necessary for cell-type specific differentiation.

Dataset Generation: Using the mouse, a well-established model of cortex development[15], we will generate multi-omic and spatial transcriptomic data at days 10, 14, and 17 of embryonic development, as this is a period characterized by migration, proliferation, neurogenesis, and dendritic development and synaptogenesis[15,37]. At each time point, we will harvest the prefrontal cortex of 6 embryos. To reduce single sample variation, the cells of three of the embryos will be pooled before being subjected to the Multiome 10x Genomics kit protocol[38]. At each timepoint, we subject two embryos to the two-dimensional STARmap protocol[1] (1020 genes), and one embryo to the three-dimensional protocol (28 genes). We choose the STARmap protocol because it is capable of providing both two and three-dimensional spatial transcriptomics datasets, and previous integrations of STARmap datasets with scRNA-seq data has provided rigorous and informative results[7].

Experiment Analysis Design: Following data generation, we can leverage the bioinformatics tools developed in Aims One and Two to evaluate the relationship between the transcriptome, the epigenome, and spatial location (**Fig. 4**). Using GRiNMF, we will integrate the spatial transcriptomic data from each sample with the multi-ome data from its corresponding time point. For each cluster, we will identify differentially expressed genes with the Wilcoxon test, and perform Gene Set Enrichment Analysis[39] to identify key developmental pathways. Similarly, we can identify differential chromatin accessibility peaks for each cluster.

We leverage the gene-level counts of the multi-omic scATAC data to predict chromatin accessibility profiles for the spatial transcriptomic cells by applying the WSNN algorithm to the integrated datasets. For quality assurance, we will compare pseudo-bulk profiles of the multi-omic scATAC with pseudo-bulk values of our predicted values using the Pearson's coefficient.

After predicting gene-level chromatin accessibility profiles for cells with spatial coordinates, we can examine how gene expression and chromatin accessibility are impacted by spatial position. To examine spatially expressed epigenomic and differentially accessible genes, we will apply SPARK (Spatial pattern Recognition via Kernels)[16], a tool that has previously been used to identify spatially expressed genes using generalized linear spatial models. We propose utilizing SPARK to identify spatially differential chromatin accessibility, a novel application of the tool, as well as spatially expressed genes. For further assessment of global differences between developmental stages, we will utilize TimeReg. We can also apply MAST[40] to derive key gene expression modules for each developmental stage.

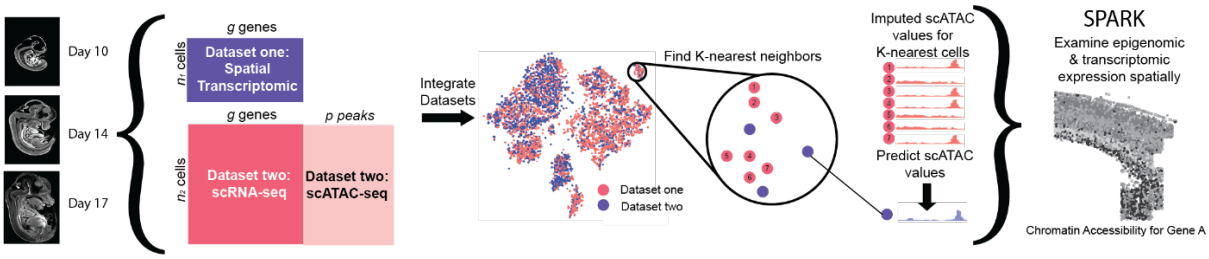

Figure 4 : At three key points in embryonic development, we will collect multiome and spatial transcriptomic data. Using GRiNMF, we will characterize the cells with spatial coordinates, and we will use WSSN to predict chromatin accessibility measures for the cells with spatial coordinates. Using SPARK, we will identify spatially expressed genes and variable regions of chromatin accessibility.

Expected outcomes: We expect the use of GRiNMF to increase the quality of dataset integration over current methods. Specifically, we would expect the clusters of resulting integrations to have higher cluster purity and greater homogeneity between the cell types placed within each cluster. We hypothesize that this will bring greater clarity when analyzing cell type-specific expression patterns and differentially accessible regions. A higher quality integration would likewise impact the application of the WSNN algorithm, as a more refined integration will correspond to the selection of more appropriate nearest neighbors.

By applying the WSSN algorithm, we derive single cell epigenetic states with single cell spatial coordinates. The pseudo-bulk expression profiles should share a high degree of correlation with pseudo-bulk multi-omic scATAC data, and we will confirm this using both the Spearman and Pearson's correlation. We expect that the application of SPARK will identify modules of both spatially expressed genes and chromatin accessibility at each time point of embryonic development. Because of the intimate relationship between transcriptome and epigenome, we expect there to be a large degree of correspondence between spatially expressed features. However, we also expect to identify regions where the spatially expressed genes and chromatin accessibility are decoupled, with modules distinct to a particular measure of cellular identity. Since epigenetic modifications typically proceed transcriptomic responses with the cell, decoupled modules could reflect the activation of distinct developmental regulatory programs. Consequently, we would focus our initial efforts on these features. Differences in chromatin accessibility between cells of the same broad annotation (i.e. interneurons) could be suggestive of fluctuations that correspond to downstream cell fate decisions, and we foresee spatially expressed chromatin values being efficacious in more clearly distinguishing cellular subtypes. Further areas of interest include identifying key driver genes that are spatially expressed, as well as those that are constitutively expressed, uncovering genes and regulatory elements that are strongly associated with a particular stage of development, and establishing those that are ubiquitous throughout development.

Potential Issues and Alternative Approaches: Although a complete published protocol is available, STARmap is not a commercialized spatial transcriptomics method. If logistical barriers prove prohibitive, another spatial transcriptomic method can be used[4,21], consequently reducing our analysis to a 2D space. Computationally, in the event that Aims 1 and 2 are unsuccessful, the current LIGER algorithm and BABEL can alternatively be used to complete the proposed Aim 3 analysis, after adjusting the quasi-likelihood equation used in SPARK to accommodate binary data[41,42]. If experimental costs exceed budget constraints, it is possible to reduce the number of spatial assays to one two-dimensional sample per development stage.

Conclusions

We present GRiNMF, a novel iNMF algorithm, that is capable of incorporating spatial information when performing dataset integration. Additionally, we develop WSSN, a shared nearest neighbor algorithm tailored to predicting scATAC-seq profiles for cells originally profiled by their transcriptome. The use of GRiNMF and WSSN, while individually informative, can be jointly utilized in an additive manner to provide predictions of scATAC-seq profiles that are likewise characterized by transcriptomic measures and spatial coordinates. Despite the flexibility of both GRiNMF and WSSN to other datasets of interest, we focus our efforts on the prefrontal cortex, as advancing the current frontier of knowledge in this area has relevance to conditions currently incurable with modern medicine, including schizophrenia, intellectual disability, ADHD, and autism disorders.

Sources

1. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, (2018).
2. Codeluppi, S. *et al.* Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* **15**, 932–935 (2018).
3. Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, (2018).
4. Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
5. Genomics, 10x. *Chromium Next GEM Single Cell Multiome ATAC + Gene Expression Reagent Kits User Guide*. (2020).
6. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
7. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873–1887.e17 (2019).
8. Cai, D., He, X., Han, J. & Huang, T. S. Graph Regularized Nonnegative Matrix Factorization for Data Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**, 1548–1560 (2011).
9. Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174**, 1015–1030.e16 (2018).
10. Moon, K. R. *et al.* PHATE: A Dimensionality Reduction Method for Visualizing Trajectory Structures in High-Dimensional Biological Data. *bioRxiv* 120378 (2017) doi:10.1101/120378.
11. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
12. Genomics, 10x. PBMCs from Healthy Donor, granulocytes removed (10K) Single Cell Multiome ATAC + Gene Expression Dataset by Cell Ranger 1.0.0. (2020).
13. Genomics, 10x. PBMCs from Healthy Donor, granulocytes removed (3K) Single Cell Multiome ATAC + Gene Expression Dataset by Cell Ranger 1.0.0. (2020).
14. Yu, Q. *et al.* Structural Development of Human Fetal and Preterm Brain Cortical Plate Based on

Population-Averaged Templates. *Cereb. Cortex* **26**, 4381–4391 (2016).

15. Chini, M. & Hanganu-Opatz, I. L. Prefrontal Cortex Development in Health and Disease: Lessons from Rodents and Humans. *Trends Neurosci.* **44**, 227–240 (2021).
16. Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* **17**, 193–200 (2020).
17. Duren, Z., Chen, X., Xin, J., Wang, Y. & Wong, W. Time course regulatory analysis based on paired expression and chromatin accessibility data. *Genome Res.* (2020) doi:10.1101/gr.257063.119.
18. Schubert, D., Martens, G. J. M. & Kolk, S. M. Molecular underpinnings of prefrontal cortex development in rodents provide insights into the etiology of neurodevelopmental disorders. *Mol. Psychiatry* **20**, 795–809 (2015).
19. Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat. Methods* **18**, 9–14 (2021).
20. Asp, M., Bergenstråhle, J. & Lundeberg, J. Spatially Resolved Transcriptomes-Next Generation Tools for Tissue Exploration. *Bioessays* **42**, e1900221 (2020).
21. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, (2015).
22. Chidester, B., Zhou, T. & Ma, J. SpiceMix: Integrative single-cell spatial modeling for inferring cell identity. *bioRxiv* 2020.11.29.383067 (2020) doi:10.1101/2020.11.29.383067.
23. Teng, H., Yuan, Y. & Bar-Joseph, Z. Cell Type Assignments for Spatial Transcriptomics Data. *bioRxiv* 2021.02.25.432887 (2021) doi:10.1101/2021.02.25.432887.
24. Thornton, C. A. *et al.* Spatially mapped single-cell chromatin accessibility. *Nat. Commun.* **12**, 1274 (2021).
25. Gao, C. *et al.* Iterative single-cell multi-omic integration using online learning. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-021-00867-x.
26. Kim, J. & Park, H. Toward Faster Nonnegative Matrix Factorization: A New Algorithm and Comparisons. in *2008 Eighth IEEE International Conference on Data Mining* 353–362 (2008).
27. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *bioRxiv* 2020.10.12.335331 (2020) doi:10.1101/2020.10.12.335331.
28. Yao, Z. *et al.* An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. *bioRxiv* 2020.02.29.970558 (2020) doi:10.1101/2020.02.29.970558.

29. Liu, L. *et al.* Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat. Commun.* **10**, 470 (2019).
30. Wu, K. E., Yost, K. E., Chang, H. Y. & Zou, J. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
31. Zhu, X. *et al.* Single-Cell Clustering Based on Shared Nearest Neighbor and Graph Partitioning. *Interdiscip. Sci.* **12**, 117–130 (2020).
32. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).
33. Li, Z., Kuppe, C., Cheng, M., Menzel, S. & Zenke, M. scOpen: chromatin-accessibility estimation of single-cell ATAC data. *BioRxiv* (2019).
34. Rocks, D. *et al.* Cell type-specific chromatin accessibility analysis in the mouse and human brain. *Epigenetics* 1–18 (2021).
35. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. doi:10.1101/2020.02.28.969931.
36. Gallegos, D. A., Chan, U., Chen, L.-F. & West, A. E. Chromatin Regulation of Neuronal Maturation and Plasticity. *Trends Neurosci.* **41**, 311–324 (2018).
37. Wong, M. D. *et al.* 4D atlas of the mouse embryo for precise morphological staging. *Development* **142**, 3583–3591 (2015).
38. Nuclei Isolation from Complex Tissues for Single Cell Multiome ATAC + Gene Expression Sequencing, Document Number CG000375 Rev B. *10x Genomics* (2021).
39. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
40. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
41. Rousset, F. & Ferdy, J.-B. Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography* **37**, 781–790 (2014).
42. Lin, P.-S. & Clayton, M. K. Analysis of binary spatial data by quasi-likelihood estimating equations. *aos* **33**, 542–555 (2005).