Pseudotemporal Analysis of Genetic Regulatory Networks in Neurodegenerative Disorders

Dhamanpreet Kaur, Sebastian Pineda, Samuel Sledzieski

December 10, 2019

Abstract

Single-cell sequencing technology has enabled the computational dissection of the genetic programming of individual cells with astonishing resolution. However, single-cell transcriptomics have been largely limited to gene expression profiling and the field of single-cell genomics as a whole lags far behind what has been accomplished at the tissue level over the last decade. Arguably, the biggest contributing actor to this is the cost and sparsity of data generated by such experiments, which can make acquiring sufficient data at enough time points difficult if not prohibitive. Pseudotime analysis has been presented as a possible solution to this problem, allowing us to reconstruct time-series data from a single-experiment by assuming that subsets of cells are in different developmental time points. We combined pseudotime analysis techniques with recently developed cell type clustering and annotation methods to identify cellular states associated with temporal disease progression and use said states as an input to a dynamic Bayesian network model. We intend to use this model to identify and dissect gene regulatory networks which are critical to understanding the molecular mechanisms of disease progression with cellular resolution, but have been largely understudied in the context of single-cell genomics. We apply this model to curated single-nucleus RNA sequencing data sets from human Alzheimer's and Huntington's disease brain samples and demonstrate that gene expression regulation differs across cell types and at different stages of disease. We are able to recover known regulatory interactions as well as identify potential new interactions across time points.

1 Introduction

Gene regulatory networks (GRNs) govern the expression of genes via the underlying relationships between transcription factors and signalling molecules. Understanding these networks holds great clinical relevance as it provides insight into effective targets for treating a disease. Substantial work has been done in using computational methods to predict these networks from large-scale transcriptome data [1, 2, 3, 4]. While this data has provided significant information pertaining to gene regulation, these bulk technologies are limited in that they take the average across all cell types in a sample [5]. Recent studies show that gene regulation is cell specific [6, 7], so a single-cell resolution approach would allow for learning of distinct regulatory networks for different cell types. Single-cell RNA sequencing (scRNA-seq) technology provides the gene expression levels of an individual cell, which can show more precise gene regulation interactions. The application of network prediction methods remains underdeveloped for the fairly recent advancement of single-cell RNA seq data, as prior studies have used approaches adapted from techniques used on bulk data.

Additionally, genetic regulatory networks have an important time element. While many methods for reconstruction of GRNs infer undirected networks, the direction and weight of these edges have a meaningful biological analogue in the regulatory impacts that genes and their protein products have on each other across time. GRN models built from dynamic processes typically involve methods such as Boolean networks [8, 9], Dynamic Bayesian Networks [1], co-expression networks [10], ordinary differential equations [11], and variational Bayesian inference

[12]. A handful of recent studies have developed approaches that first order the cells along a temporal trajectory, using time or pseudotime data to infer GRNs [12, 13]. However, these applications have each used samples explicitly taken at different time points, or have used pseudotime analyses from the area of developmental biology.

This study seeks to explore the application of Dynamic Bayesian networks (DBNs), which have been used with bulk data [14], to infer patterns of gene regulation and gene expression in neurodegenerative diseases. By using DBNs, it is possible to model the inherently time-dependent regulatory interactions between genes. We will develop a novel pseudotime ordering framework to train a DBN using scRNA-seq data from fully developed brain samples, and apply this framework to data sets from Alzheimer's Disease (AD) and Huntington's Disease (HD) in an attempt to elucidate the regulatory mechanisms behind neurodegenerative disease.

Focusing on AD and HD provides us with the opportunity to study GRNs linked to neurodegeneration. As a well-studied monogenic disease, many of the GRNs perturbed in HD have been characterized, at least at the bulk level, whereas the same cannot be said of AD, a polygenic disease where the pathogenic cause is unknown, where these same mechanisms remain elusive. Despite these differences, it is known that this diseases share many similarities at the molecular level and the ultimately lead to the same fatal consequence. Thus our approach is applicable to the study of both conditions, and elucidating the regulatory perturbations of one disease will enable us to shed light on the same or similar molecular mechanisms in the other.

2 Methods

2.1 Data Preprocessing

Barcoded and indexed sequencing libraries are demultiplexed and aligned to a pre-mRNA reference genome. Instances of each gene with a unique molecular identifier and barcode are counted to produce a sparse feature-barcode matrix. Each matrix contains cells where each column is a separate cell and the rows represent the levels of gene expression. This matrix is batch corrected using Harmony and reduced and normalized using ACTIONet.

2.2 Annotation of Cell Types

We create an ACTIONet [18] from the reduced and batch corrected expression data to discover archetypes within the data set. We then use a curated combination of two sets of published pre-frontal cortex markers from Velmeshev et al. [15] and the PsychENCODE Consortium [16] to annotate each cell as one of the cell types. We subset the cells annotated as each cell type and created a sub-ACTIONet for each of the cell types. Separately, we used a set of Alzheimer's Disease phenotype genetic markers from the NanoString database and found the archetype which was most enriched for disease association. We decided to focus our study first on microglia based on the level of enrichment for the disease associated archetype and the quality of the data.

2.3 Pseudotime Ordering

2.3.1 Identification of Disease Archetype

Within each cell type sub-ACTIONet, we again use the NanoString AD markers to find the cell state which is maximally enriched for disease association. When looking at the microglial cells, we also considered enrichment for activated microglia as a determining factor for the associated state. We additionally used the Braak stage, final cognitive diagnosis, and CERAD score for each individual as supplementary information in selecting the disease archetype. However, these metrics are subjective and are only at the resolution of an individual patient, which means that heterogeneous cells in an individual would all be labeled the same. Because of this, we leaned more on the disease markers for identifying the correlated archetype.

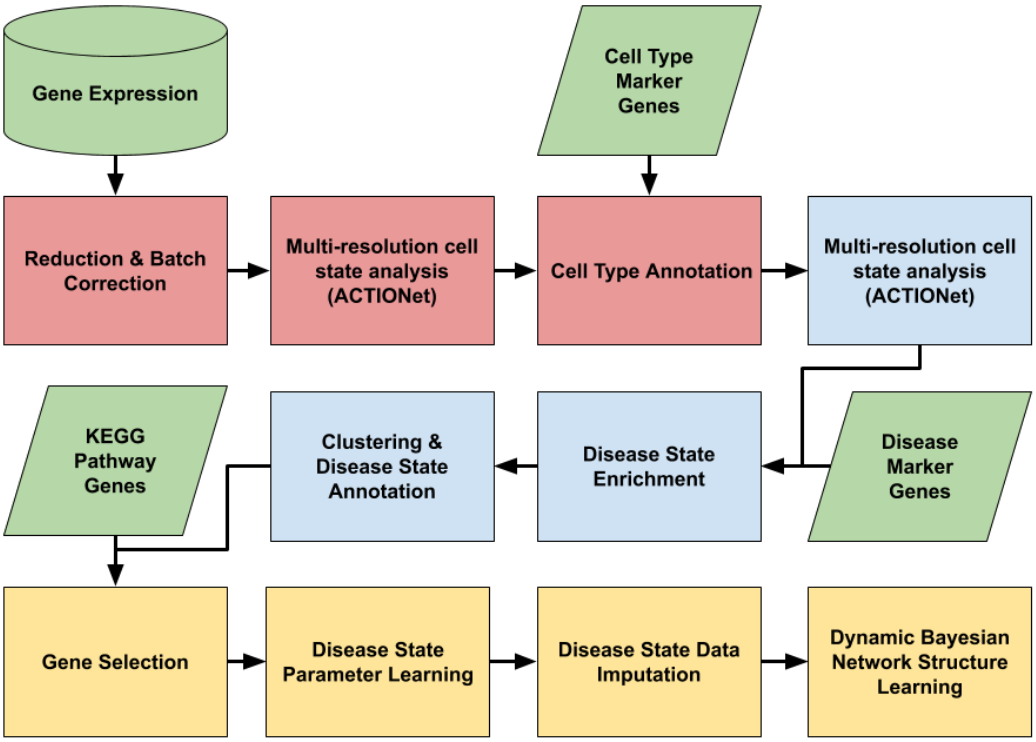


Figure 1: Our analysis starts with the raw counts of gene expression. Using ACTIONet and incorporating marker genes for brain cell types from [15] and [16], we perform cell type annotation (red boxes). We then subset the microglia cells, and incorporate disease marker genes from the Nanostring database to identify cell states correlated with disease. This cell state is used to assign a disease score to each cell, which is then used to cluster the cells and order them according to pseudotime (blue boxes). Finally, we use disease pathway genes from the KEGG database [17] to filter our library, model the data and use the model to simulate observations across all time points, and learn the structure of a Dynamic Bayesian Network (yellow boxes). These networks are analyzed to identify import regulatory elements.

2.3.2 Ordering and Clustering of Cells

The selected archetype is then used to assign scores to each cell. This score corresponds to the normalized loadings for each cell with the core archetype which is correlated. The cells can then be sorted by this score, and clustered into stages of disease progression. As a first naive clustering, we simply discretized the disease score into four bins of equal width, corresponding to the four core archetypes identified. Additionally, we tried binning into different numbers of bins of equal size, as well as k-means clustering of the cells using the expression values of the top 100 differentially expressed genes that define the disease associated archetype. The final clustering used was a simple binning of disease scores (see Figure 3C).

2.3.3 Pseudotime Ordering using Monocle 3

We also attempted to use Monocle 3 [19] to construct a pseudotemporal ordering of our cells. However, the clustering and trajectories inferred by Monocle made little biological sense in the context of our data. As a result, we proceeded with the ordering learned from the ACTIONet disease score. See Section 3.3 for more detail.

2.4 Learning Dynamic Bayesian Network structure

2.4.1 Gene Selection

To limit computational complexity and reduce the impact of noise, it was necessary to reduce the number of genes used in our network from the total library size (17,926 in our AD assay). To focus on biologically relevant genes, we first filtered our library down using the KEGG Alzheimer's Gene set [17, 20]. This resulted in a set of 139 genes. We plotted the average expression level of each of these gene at each pseudotime point to visualize how gene expression varied across pseudotime, and clustered the genes (See FigureS3A). We selected the top 30 rows of this clustering, which appeared to be the most variable genes that would provide the most information for network learning while reducing noise. We used this set of 30 genes as the variables for our analysis, and since we had 4 pseudotime points this resulted in 120 variables in our dynamic Bayesian network.

2.4.2 Data Imputation

The ideal input to learning the DBN would be the expression level of each gene in a singular cell over the course of disease progression. However, because of the limitations of single-cell RNA sequencing technology, we only have a snapshot of the gene expression levels of a cell at a singular time point. As a result, we use the data within each cluster to simulate the gene expression levels of a typical cell at that stage of the disease. There are multiple approaches that can be used for data imputation. For instance, we can sample from a statistical distribution – such as a normal or Poisson distribution – using parameters learned from each gene's expression levels within a given disease stage. Another implementation uses the Splatter package [21] which employs a more sophisticated method of learning a gamma-Poisson distribution and can simulate additional single-cell specific features such as dropout. The drawbacks of these approaches is that they assume the gene expression levels within a disease stage are independent. One alternative to this is to use bootstrapping to sample from cells at each disease stage, but the imbalance in cluster sizing may bias the variance of the simulated data. Another option which models intra-time point dependency is employing a static Bayesian network to learn the relationships between gene expression levels within cells at the same stage in disease progression. Using either the distributions or the learned static networks, we simulate n = 5,000 data points representing gene expression at each stage and concatenate these into one large matrix.

2.4.3 Learning the DBN

We then learn a Dynamic Bayesian network from this data by only allowing edges from a gene at time t_i to a gene at time t_k if $i < k$. Bayesian networks rely on the Markov property that the value of any random variable only depends on its parents. We experiment with a variety of constraint-based algorithms – such as Grow-Shrink, Incremental Association, Hiton Parents and Children – as well as score-based algorithms – such as Hill Climbing and Tabu Search – via the "bnlearn" package in R to construct the directed acyclic graphs from the data. [22] Constraint-based models use an inductive causation model that first learns the structure of undirected edges in the graph using conditional independence tests and then computes parameters that satisfy the structure. These models offer a causal interpretation and are consequently useful for inferring gene activation or inhibition. Score-based models assign a score to each candidate Bayesian network and try to maximize it with some heuristic search algorithm. These are not commonly used for gene regulatory networks but we include them for the sake

of comparison in model performance. At this stage of the project, the choice to use one learning method over another was largely based on the number and consistency of edges learned, but more robust feedback from network validation will be used in the future to better refine this choice.

2.4.4 Identification of Consistent Network Edges

Due to the step of data imputation, the edges that are learned in the Dynamic Bayesian Network are fairly stochastic. To overcome this, we use a bootstrapping approach. We repeat the process of imputing data and learning the DBN several times, and then identify the strength of the edges by the frequency with which they appear over all of the learned networks. In this study, we simulated 100 different data sets using the learned parameters, learned a network at significance level $\alpha = 0.05$ for each replicate, and combined the results. Figure 4 shows the learned edges and the number of bootstrap replicates in which they occurred (with edges occurring fewer than a bootstrap threshold of 8 set to 0 for clarity).

2.5 Exploration and Validation of Learned Network

We transform the bnlearn representation of the network into an adjacency matrix and calculate the node degrees (number of parents and number of children since our network is a directed acyclic graph), betweenness centrality, and closeness centrality for each nodes. We plot the adjacency matrix sorted and divided into layers to visualize the arcs in the network. Additionally, we plot the number of children for each node to easily identify genes which many other nodes are dependent on. We compare high-degree and highly central nodes and interactions to the literature and to other methods for network learning to assess the validity of learned interactions.

2.6 Network Learning with GRNVBEM

After selecting our set of 30 genes, we used the ACTIONet disease score to construct a complete pseudotime ordering of all cells, which we used as input for GRNVBEM [12]. The tool runs in MATLAB and requires only the ordered count matrix as input, which made it easy to run and compare with our own methods. However, our method differs from GRNVBEM in that we learn relationships between gene expression levels at different time points, while GRNVBEM only learns a single network. We compare the network learned by GRNVBEM with the dynamic network learned using our method in Section 3.6.

3 Results

3.1 Multi-resolution Analysis of Single-Cell Transcriptomic Data Allows for Cell Type Annotation

Using ACTIONet, we were able to clearly identify cell types in the Alzheimer's cells from the pre-frontal cortex, as seen in Figure S1A. We show in Figure S1B that archetype 5 has the strongest association with disease across all cell types. Archetypes are distinct profiles of gene expression, and the top 5 DEGs which define archetype 5 are PLXDC2, C10orf11, DOCK4, ARHGAP254, and SLC1A3. This does not mean that those genes are markers for disease, but rather that they best separate cells which are correlated with disease from cells in other states.

While the approach could be applied to any of the cell types, for Alzheimer's disease we focused on the microglia in this study. This choice was made because of the significantly increased enrichment for disease association in microglia, as seen in Figure S1C. Since our aim is to study disease progression, we a cell type where cells are clearly associated with disease. Further, when we color all cells by their level of disease association, there is a range of enrichment across microglia cells which supports the need to perform a single-cell resolution study of regulatory networks, especially in complex disease (see Figures 3A and S1D).

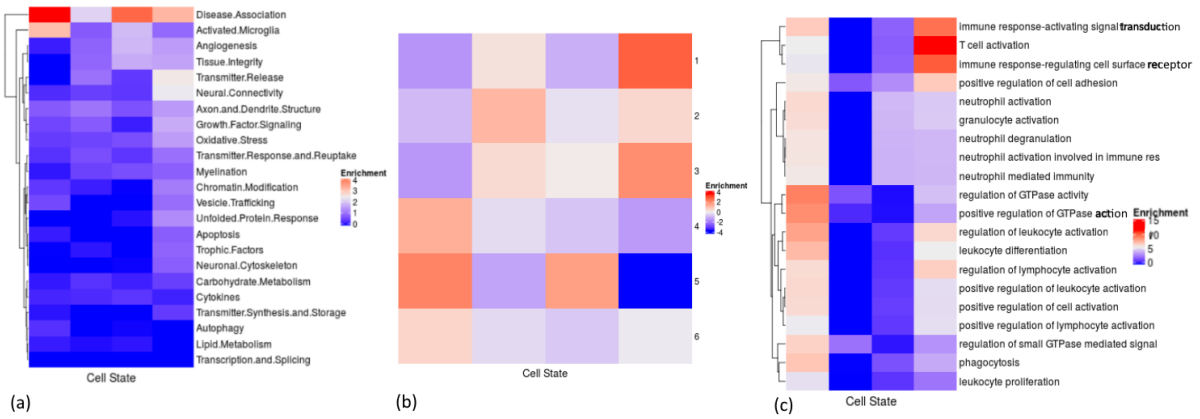


Figure 2: (a) Cell state 1 is most strongly enriched for disease association and activated microglia. (b) The same cell state is more strongly enriched in the later Braak stages, and negatively enriched in the earlier Braak stages. (c) There is a positive enrichment for GTPase and leukocyte activation in the disease associated state. Archetype four, which is strongly diminished in later Braak stages has a large enrichment for immune response pathways.

3.2 Within-Cell-Type Heterogeneity Correlates with Disease

Using the sub-ACTIONet constructed only on the microglial cells, we are able to identify the first cell state as the most strongly enriched for disease association (Figure 2A). This is supported in Figure 2B, where the first state is the most strongly enriched for the late Braak stages. It is interesting to note also that the fourth cell state, which is enriched in early Braak stages and diminished in later Braak stages, is enriched for a number of immune pathways. Recent work has shown that Alzheimer's Disease, often traditionally thought of as a neurological disease, has a significant immune component [23]. In the early stages of disease, there is strong enrichment for immune and T cell activation pathways, but in later stages of disease progression the most differentially expressed genes correspond to positive regulation of GTPase and leukocyte activation.

We verify that the gradient of disease enrichment observed corresponds to distinct sub-populations in the microglia. Using ACTIONet, we annotate cells into one of the four cell states, as seen in Figure 3B. These cell state annotations correspond strongly with the gradient of enrichment for disease progression. We assigned each cell to one of four pseudotime layers, which closely represents the density of disease scores and represents the diversity of disease progression among microglial cells (Figure 3C). We further show that a pseudotime model is necessary since disease progression enrichment explains a significant portion of the variance in gene expression (Figure 3D). The cell state clusters determined here were used for learning the Dynamic Bayesian Network.

3.3 Disease State Enrichment Tracks Improves Progression Tracking

The most widely used software currently used for pseudotime and trajectory inference is Monocle [19]. Monocle has primarily been applied to embryonic development and stem cell differentiation, but we were not able to find any instances of its application for disease progression tracing. When we applied Monocle to our data, the resulting clustering and trajectory did not make sense biologically, likely as consequence of the packages' data preprocessing pipeline which was designed to operate on dense, monoclonal stem cell RNA-seq data and was incompatible with our AD and HD data. (see Figure S2). However, using our method of disease state enrichment using ACTIONet, we were able to get consistent clustering and ordering of cells which corresponded with biological pathway enrichment as well as clinical metadata (see Figures 2, 3).

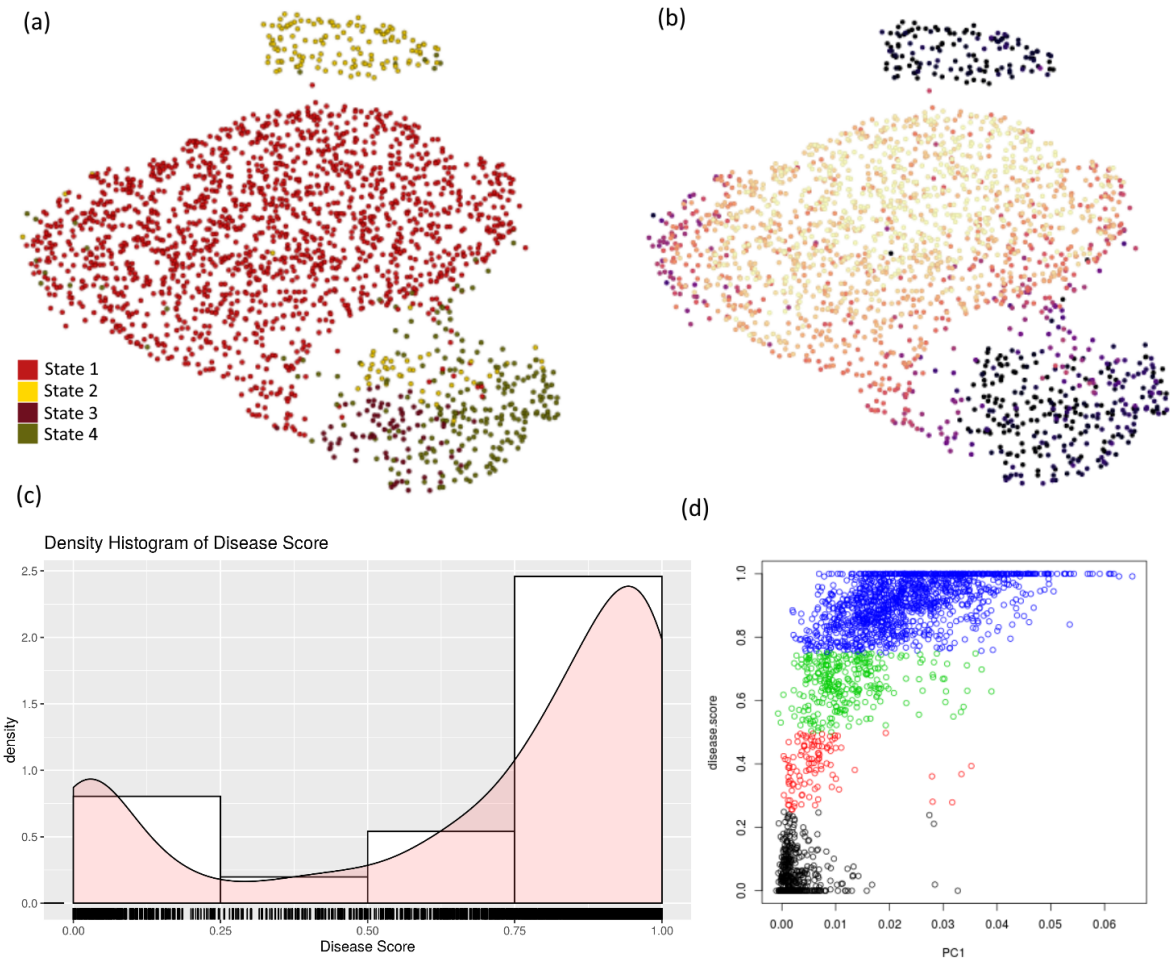


Figure 3: (a) There are distinct sub-populations of disease progression enrichment among the microglial cells. (b) ACTIONet annotation of cells into one of the four core cell archetypes, where cells in state 1 are most correlated with late disease progression. (c) The disease scores are well distributed and can be binned into four clusters. Four clusters were chosen to correspond to the four core archetypes discovered by ACTIONet. (d) The disease enrichment score of cells explains the variance in gene expression of microglial cells.

3.4 Imputation of Cell State Observations Allows for Dynamic Learning

Current technology does not allow the monitoring of gene expression levels in a singular cell over the course of disease progression. However, we can impute parameters for gene expression at each time point and simulate observations to increase the size and robustness of our data. We simulated new observations in two ways. First, we learned a static Bayesian network with gene expression levels across cells in the same disease stage. We also fit a Poisson distribution to the levels of gene expression at a disease stage. Both of these methods allow us to model the temporal interactions of genes without imposing the loss of individual cell specificity seen in bulk RNA sequencing technologies.

We verified that our simulated data was realistic by plotting the variation of each gene across our time steps. Using both the Bayesian network simulation and Poisson distribution simulation, our selected genes vary in the same way and cluster similarly to the true gene expression values with the raw count data (see Figure S3). This imputation also allows us to generate several new data sets to perform replicate network inferences (see Section

2.4.4). With these methods combined, we were able to recover edges which occurred significantly more often than they would due to chance, and calculate strengths for each edge learned this way (see Figure S5).

3.5 Gene Expression Dependence Across Pseudotime Recovers Known Interactions

Several constraint-based algorithms for the graphs have been able to detect edges indicating temporal relationships for gene activation and inhibition. The learning process runs in less than 30 seconds across 30 genes with 5,000 generated samples. The best constraint-based algorithms show that about 8-12% of edges are captured more than once in one hundred iterations of network learning (See Figure S5 for full results on consistency of edge detection across various network learning methods). An adjacency matrix for the learned network, along with a Cytoscape plot of the network can be found in Figure 4.

An analysis of our learned Dynamic Bayesian Network using Poisson simulation reveals a number of regulatory interactions and significant genes well known from the literature. The genes with the largest number of children (expression values dependent upon them) are APP, APOE, and GAPDH in the first time step (see Figure 4A. These genes are all well known to have a significant impact in the development of Alzheimer's disease [24, 25, 26]. Additional high degree genes include TNFRSF1A (Identified in [27]), PLCB4, and NDUF4 in the first time step. While these results are consistent with the literature, the fact that all top degree genes are from the first time step is likely due in part to the fact that there are more available arcs to variables in the first layer.

To combat this, we also looked at closeness and betweenness of each node in the graph, which should give a better sense of which genes are central in the graph, and therefore potentially important in understanding gene regulation in Alzheimer's disease. Closeness centrality tends to represent the ability to rapidly spread information throughout a network, and betweenness centrality tends to represent the ability to control information flow throughout a network. This analysis supported the insight gained by looking at degree centrality, as well as provided new information. The top genes by closeness centrality were ADAM10, ADAM17, APAF, APOE, APP, and ATF6 across several time points. Of these, there is strong support in the literature for the importance of the two ADAM genes, APOE, APP, and ATF6 [28, 29, 25, 24, 30].

The top genes by betweenness centrality were CACNA1D at the second time point, ADAM10 at the third time point, ADAM17 at the first and second time point, and PLCB4 at the first time point. This is consistent with the importance of the ADAM genes reported in the literature and indicated by closeness centrality. PLCB4 appears as both a high degree and high centrality node, but appears in only a single place in the literature [31]. The role of PLCB4 in Alzheimer's may warrant further study. ITPR2 is also an interesting gene to be identified as highly central because of its appearance in the GRNVBEM network, which is described in more detail in Section 3.6.

3.6 Dynamic Approach Learns Important Interactions Not Recovered By a Static Approach

Previous attempts at learning gene regulatory networks from single-cell RNA sequencing data have used pseudotime ordering, but do not show the temporal nature of the gene interactions. The approach developed here indicates which genes are activated or inhibited at specific stages of disease progression. This could hold valuable clinical insight for developing drugs or treatments that specifically target cells to prevent them from further deterioration. In contrast, tools such as GRNVBEM which use pseudo-temporally sorted data report a single network rather than reporting the changes in regulatory activity over time. Our approach recovers the edge from ATF6 to ITPR2 which have significant weight in the GRNVBEM network (see Figures 4 and S4). We identify ITPR2 as a gene with high betweenness centrality, and it is one of the most central nodes in the GRNVBEM network. However, our approach recovers more significant interactions for ITPR1, with dependencies on itself, ITPR2, and PSEN1 (among others) at various different time points. Additionally, GRNVBEM does not suggest the regulatory importance of APOE, ATP, or GAPDH, which are all well known to play a significant role in disease

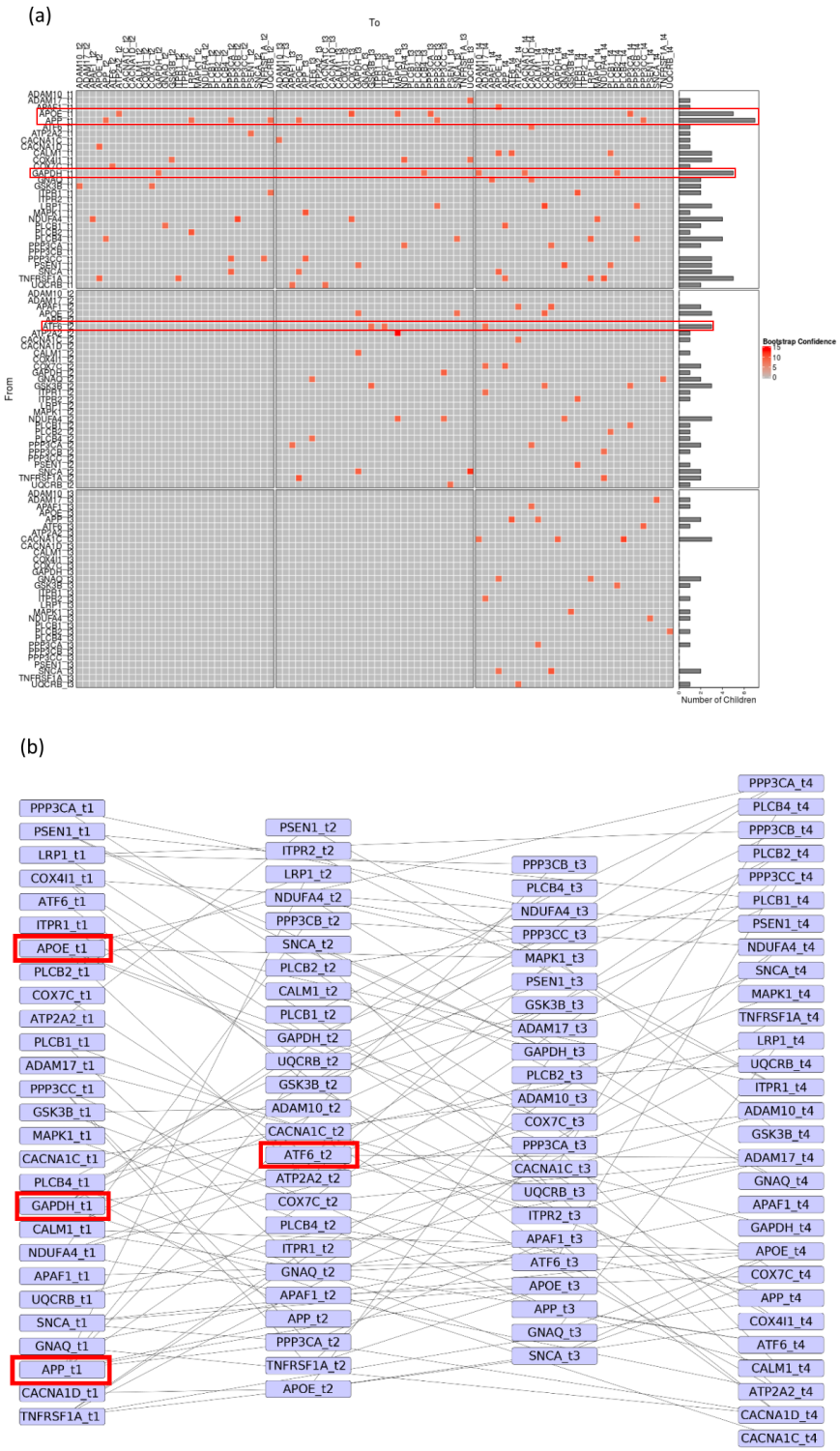


Figure 4: (a) Dynamic Bayesian Network generated by Incremental Association Method, with 100 bootstrap replicates of 5,000 samples simulated from learned Poisson distributions. For clarify of visualization, edges which appear fewer than 8 times are removed. APOE, APP, GAPDH, and ATF6, which have all appeared in the literature as significant regulatory elements, are highlighted. (b) Layout of network from (a) using Cytoscape. The same elements are highlighted. Edges which appear fewer than 8 times do not appear in this layout.

progression. This suggests that a continued study of the dynamics of gene regulatory networks could play an important role in identifying key regulatory elements.

3.7 Dynamic Network Learning Applied to Huntington's Disease

We applied a modified implementation of the aforementioned approach to out scRNA-seq data set indirect pathway spiny projection neurons (iSPN) from human Huntington's disease caudates. The counts matrix was batch-corrected and log-normalized using the fastMNN algorithm. In order to study pseudotime trajectories and the temporal relationship between dysregulated genes in a pathway-specific manner, we used only counts corresponding to genes in the Huntington's disease KEGG pathway as input to ACTIONet. The resulting cell distribution and cluster assignments are shown in Figure 5. The spatial positioning of cells is primarily driven by *HTT*, whose expression, and in particular its down-regulation, is a canonical marker of HD progression. The gene expression heat map in Figure 5 recapitulates the findings of previous single-cell studies of HD, namely that late stage HD is characterized by extreme cytosolic up-regulation of mitochondrial-encoded mRNAs. The adjacency matrix generated by our DBN shows that the node with the most children corresponds to UQCRC2, a gene encoding a sub-unit of electron transport chain Complex 3, whose down-regulation has been previously shown to be marker of mitochondrial dysfunction and characteristic of disease progression.

4 Discussion

4.1 Summary

Our preliminary results seem promising and validate the approach of using pseudotime and Dynamic Bayesian Networks to perform a study of dynamic regulatory networks in different cell types. There are clearly cell types where gene expression is more correlated with disease than others, and within the cell type there is further heterogeneity which validates looking further at the state of the cells, and how they change as disease progresses. We were able to successfully identify several well-known genes as significant genetic regulators of disease, and to develop new insight into how regulation changes across time. Our approach has been applied to cell types in both Alzheimer's Disease and Huntington's Disease, and is broadly applicable to several cell types and neurodegenerative diseases, and potentially other progressive diseases.

4.2 Future Research

While this project has made substantial progress in developing a methodology for learning dynamic gene interactions over the course of disease progression, a large amount of work remains to be done in validating these networks and interpreting their biological significance. It would be extremely useful to develop a metric that reflects the validity of the network so that it could be used as a feedback mechanism for optimizing the graph learning methods. Additionally, the process of pseudotime ordering and clustering is still largely manual and requires some amount of domain knowledge. An automated clustering and ordering of cells (without identifying a disease archetype) would improve the re-usability of our methods.

4.3 Limitations

There are many variables and assumptions made throughout the process of transforming the data so that it can be captured by a Dynamic Bayesian network. The temporal nature of the model is incredibly useful and holds important clinical implications, but this first requires robust validation of the disease state clustering and data imputation steps. Since we only have a snapshot of a cell in a single stage, we are left imputing observations of gene expression levels for the same cell across multiple time points.

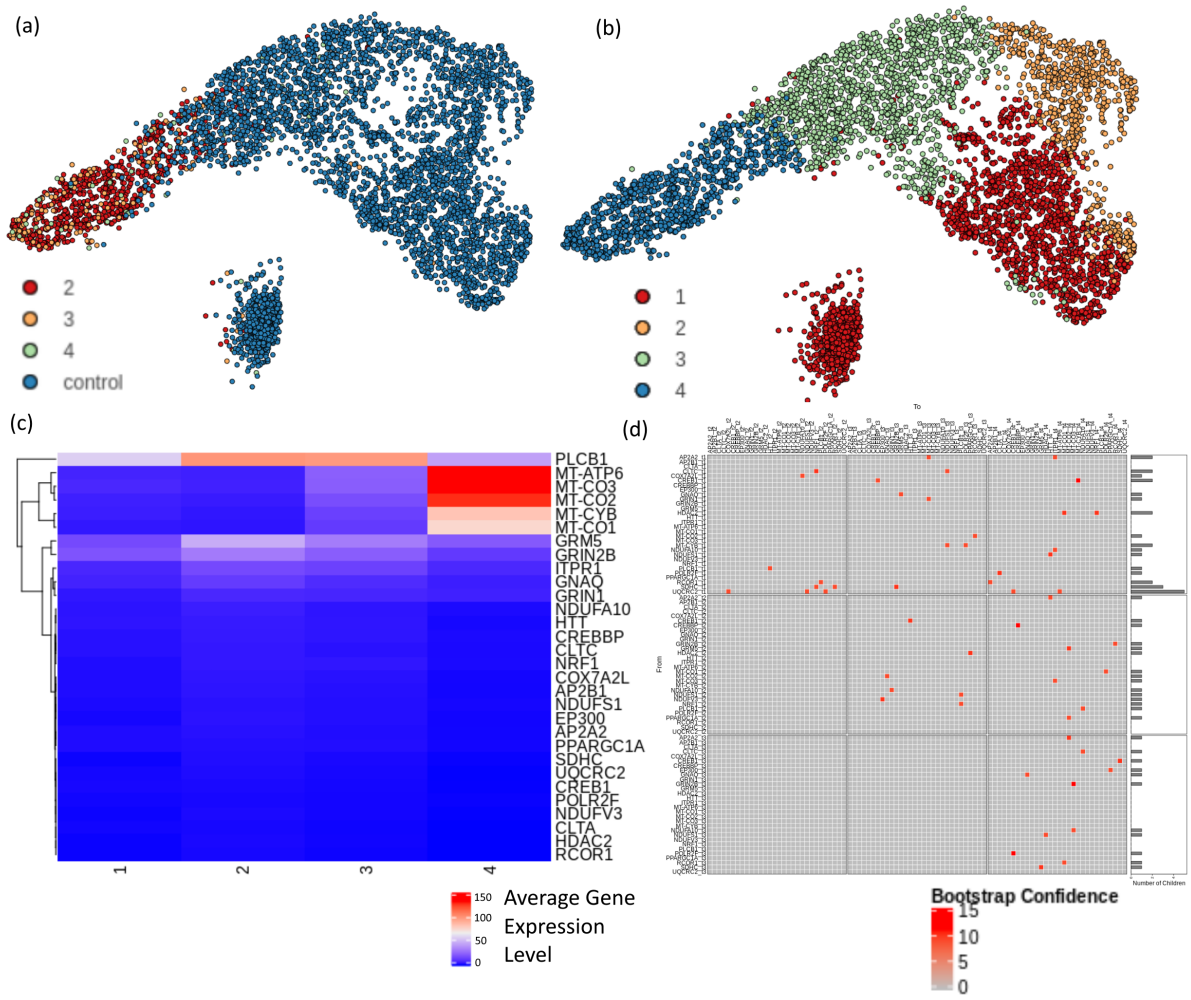


Figure 5: (a) ACTIONet clustering of iSPNs annotated by pathological grade of the patient of origin. (b) The same population of cells annotated by ACTIONet sub-cluster. (c) Gene expression heat map iSPN sub-clusters. (d) Adjacency matrix of DBN generated from iSPN gene expression counts.

5 Contribution

Sledzieski, Pineda, and Kaur jointly devised the research, performed the literature review, and contributed to writing and presentation. Sledzieski performed the cell type annotation, disease archetype analysis, pseudotime ordering, and clustering for the Alzheimer's data, contributed to the data imputation for Bayesian network learning, and contributed to the network analysis. Kaur developed the framework of learning a Dynamic Bayesian network from single-cell RNA sequencing data that has been organized into clusters ordered based on disease progression, which involves imputing data through the use of either static Bayesian network learning or bootstrapping and then proceeds to learn a directed acyclic graph between genes at various stages of the disease. Kaur also applied the GRNVBEM method to the pseudotime ordering of this data for validation and comparison purposes. Pineda performed the cell type annotation, pseudotime ordering, and clustering for the Huntington's data, applied Monocle to the Alzheimer's and Huntington's data, and contributed to the network analysis.

6 Commentary

The overall project was certainly multi-faceted and challenging. Each of the three major aims – pseudotime ordering, network learning, and biological interpretation – could individually entail significantly more depth of research. We were satisfied in that we were able to achieve promising results applying the envisioned process to two different neurodegenerative diseases. We believe this a largely novel approach for learning gene regulatory networks from single-cell RNA sequencing data, but there remains substantial work to be done in results validation before we can pursue publication. If we were to do things differently, we would use a data set from a disease with a known GRN and first determine if the method is capable of detecting the edges in that network. This would have allowed us to confirm that the assumptions made in pseudotime ordering, data imputation, and the use of a DBN structure are justified and form an accurate model for the system.

References

[1] P. Li, C. Zhang, E. J. Perkins, P. Gong, and Y. Deng, "Comparison of probabilistic boolean network and dynamic bayesian network approaches for inferring gene regulatory networks," in *BMC Bioinformatics*, vol. 8, p. S13, BioMed Central, 2007.

[2] N. A. Barker, C. J. Myers, and H. Kuwahara, "Learning genetic regulatory network connectivity from time series data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 1, pp. 152–165, 2009.

[3] F. Liu, S.-W. Zhang, W.-F. Guo, Z.-G. Wei, and L. Chen, "Inference of gene regulatory network based on local bayesian networks," *PLoS computational biology*, vol. 12, no. 8, p. e1005024, 2016.

[4] B. Yang, W. Zhang, and J. Lv, "A new supervised learning for gene regulatory network inference with novel filtering method.," *International Journal of Performability Engineering*, vol. 14, no. 5, 2018.

[5] M. W. E. J. Fiers, L. Minnoye, S. Aibar, C. Bravo González-Blas, Z. Kalender Atak, and S. Aerts, "Mapping gene regulatory networks from single-cell omics data," *Briefings in Functional Genomics*, vol. 17, pp. 246–254, 01 2018.

[6] A. R. Sonawane, J. Platig, M. Fagny, C.-Y. Chen, J. N. Paulson, C. M. Lopes-Ramos, D. L. DeMeo, J. Quackenbush, K. Glass, and M. L. Kuijjer, "Understanding tissue-specific gene regulation," *Cell reports*, vol. 21, no. 4, pp. 1077–1088, 2017.

[7] C. S. Greene, A. Krishnan, A. K. Wong, E. Ricciotti, R. A. Zelaya, D. S. Himmelstein, R. Zhang, B. M. Hartmann, E. Zaslavsky, S. C. Sealfon, *et al.*, "Understanding multicellular function and disease with human tissue-specific networks," *Nature genetics*, vol. 47, no. 6, p. 569, 2015.

[8] M. Imani and U. Braga-Neto, "Optimal gene regulatory network inference using the boolean kalman filter and multiple model adaptive estimation," in *2015 49th Asilomar Conference on Signals, Systems and Computers*, pp. 423–427, Nov 2015.

[9] C. Y. Lim, H. Wang, S. Woodhouse, N. Piterman, L. Wernisch, J. Fisher, and B. Gottgens, "BTR: training asynchronous boolean models using single-cell expression data," *BMC Bioinformatics*, vol. 17, 2016.

[10] A. T. Specht and J. Li, "Leap: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering," *Bioinformatics*, vol. 33, no. 5, 2016.

[11] H. Matsumoto, H. Kiryu, C. Furusawa, M. S. H. Ko, S. B. H. Ko, N. Gouda, T. Hayashi, and I. Nikaido, "SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation," *Bioinformatics*, vol. 33, pp. 2314–2321, 04 2017.

[12] M. Sanchez-Castillo, D. Blanco, I. M. Tienda-Luna, M. C. Carrion, and Y. Huang, "A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data," *Bioinformatics*, vol. 34, pp. 964–970, 09 2017.

[13] L. F. Iglesias-Martinez, W. Kolch, and T. Santra, "BGRMI: A method for inferring gene regulatory networks from time-course gene expression data and its application in breast cancer research," *Scientific reports*, vol. 6, p. 37140, 2016.

[14] B. Yu, J.-M. Zu, S. Li, C. Chen, C. Rui-Xin, W. Wang, Y. Zhang, and M.-H. Wang, "Inference of time-delayed gene regulatory networks based on dynamic bayesian network hybrid learning method," *Oncotarget*, vol. 8, no. 46, 2017.

[15] D. Velmeshev, L. Schirmer, D. Jung, M. Haeussler, Y. Perez, S. Mayer, A. Bhaduri, N. Goyal, D. H. Rowitch, and A. R. Kriegstein, "Single-cell genomics identifies cell type–specific molecular changes in autism," *Science*, vol. 364, no. 6441, pp. 685–689, 2019.

[16] D. Wang, S. Liu, J. Warrell, H. Won, X. Shi, F. C. Navarro, D. Clarke, M. Gu, P. Emani, Y. T. Yang, *et al.*, "Comprehensive functional genomic resource and integrative model for the human brain," *Science*, vol. 362, no. 6420, p. eaat8464, 2018.

[17] M. Kanehisa, Y. Sato, M. Furumichi, K. Morishima, and M. Tanabe, "New approach for understanding genome variations in kegg," *Nucleic acids research*, vol. 47, no. D1, pp. D590–D595, 2018.

[18] S. Mohammadi, J. Davila-Velderrain, and M. Kellis, "A multiresolution framework to characterize single-cell state landscapes," *bioRxiv*, 2019.

[19] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. Pliner, and C. Trapnell, "Reverse graph embedding resolves complex single-cell developmental trajectories.," *BioRxiv*, 2017.

[20] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545–15550, 2005.

[21] L. Zappia, B. Phipson, and A. Oshlack, "Splatter: simulation of single-cell rna sequencing data," *Genome biology*, vol. 18, no. 1, p. 174, 2017.

[22] M. Scutari, "Learning bayesian networks with the bnlearn r package," *Journal of Statistical Software*, vol. 35, 2010.

[23] H. Mathys, J. Davila-Velderrain, Z. Peng, F. Gao, S. Mohammadi, J. Z. Young, M. Menon, L. He, F. Abdurrob, X. Jiang, *et al.*, "Single-cell transcriptomic analysis of alzheimer's disease," *Nature*, p. 1, 2019.

[24] R. J. O'Brien and P. C. Wong, "Amyloid precursor protein processing and alzheimer's disease," *Annual review of neuroscience*, vol. 34, pp. 185–204, 2011.

[25] C.-C. Liu, T. Kanekiyo, H. Xu, and G. Bu, "Apolipoprotein e and alzheimer disease: risk, mechanisms and therapy," *Nature Reviews Neurology*, vol. 9, no. 2, p. 106, 2013.

[26] D. A. Butterfield, S. S. Hardas, and M. L. B. Lange, "Oxidatively modified glyceraldehyde-3-phosphate dehydrogenase (gapdh) and alzheimer's disease: many pathways to neurodegeneration," *Journal of Alzheimer's Disease*, vol. 20, no. 2, pp. 369–393, 2010.

[27] F. E. McAlpine and M. G. Tansey, "Neuroinflammation and tumor necrosis factor signaling in the pathophysiology of alzheimer's disease," *Journal of inflammation research*, vol. 1, p. 29, 2008.

[28] X.-Z. Yuan, S. Sun, C.-C. Tan, J.-T. Yu, and L. Tan, "The role of adam10 in alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 58, no. 2, pp. 303–322, 2017.

[29] D. Hartl, P. May, W. Gu, M. Mayhaus, S. Pichler, C. Spaniol, E. Glaab, D. R. Bobbili, P. Antony, S. Koegelsberger, *et al.*, "A rare loss-of-function variant of adam17 is associated with late-onset familial alzheimer disease," *Molecular psychiatry*, p. 1, 2018.

[30] A. Salminen, A. Kauppinen, T. Suuronen, K. Kaarniranta, and J. Ojala, "Er stress in alzheimer's disease: a novel neuronal trigger for inflammation and alzheimer's pathology," *Journal of neuroinflammation*, vol. 6, no. 1, p. 41, 2009.

[31] Y. R. Yang, D.-S. Kang, C. Lee, H. Seok, M. Y. Follo, L. Cocco, and P.-G. Suh, "Primary phospholipase c and brain disorders," *Advances in biological regulation*, vol. 61, pp. 80–85, 2016.

Supplements

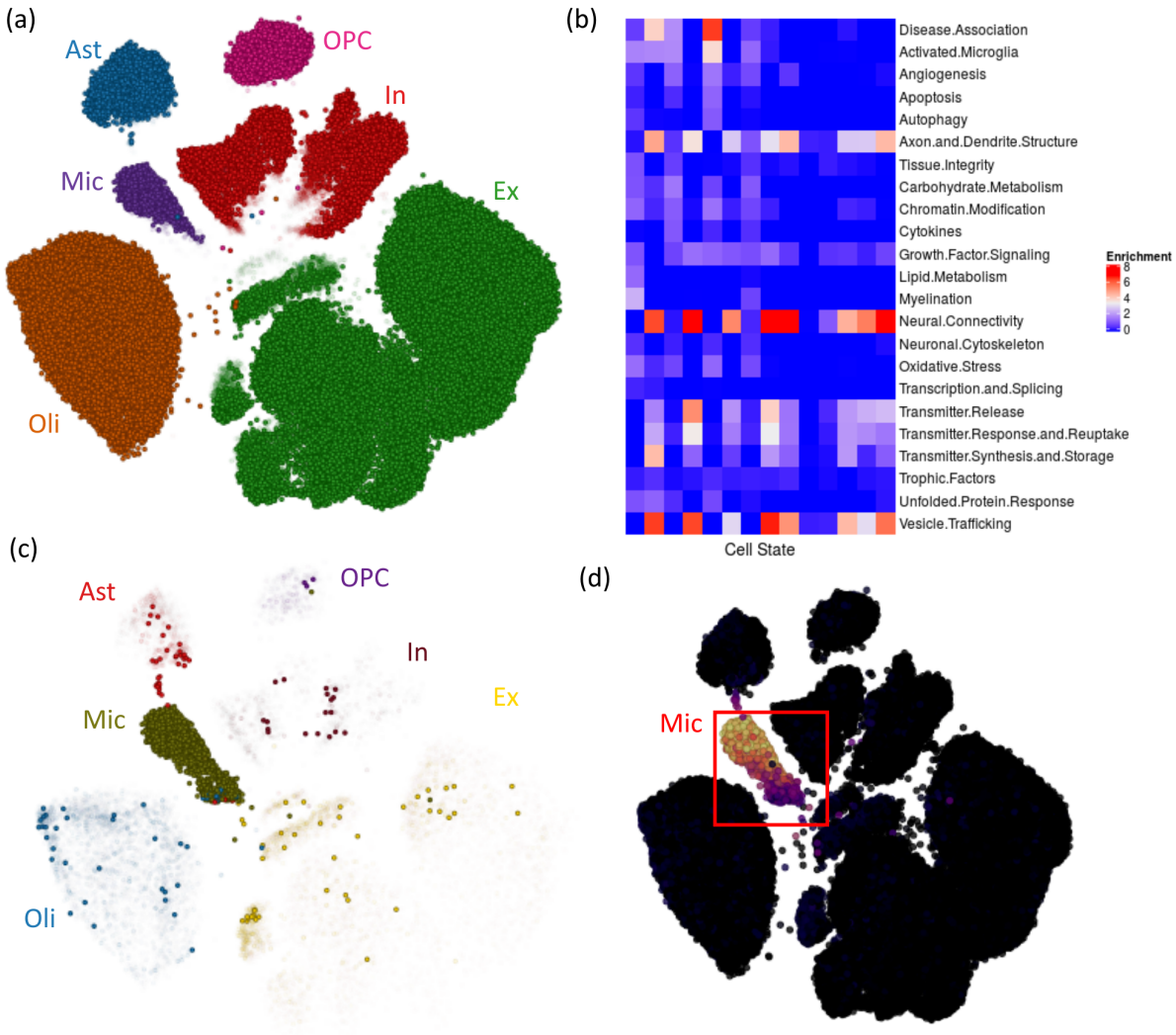


Figure S1: (a) ACTIONet layout of cells colored by cell type. Cell types with lower confidence of type assignment have higher transparency. (b) The fifth archetype has the strongest enrichment for disease association. (c) Cell types are plotted again with the transparency as the level of enrichment for the fifth archetype. Microglia have the strongest enrichment for the disease archetype. (d) There is a gradient of disease progression association within the microglial cells, which validates the pseudotime study of intra-cell type heterogeneity and disease progression.

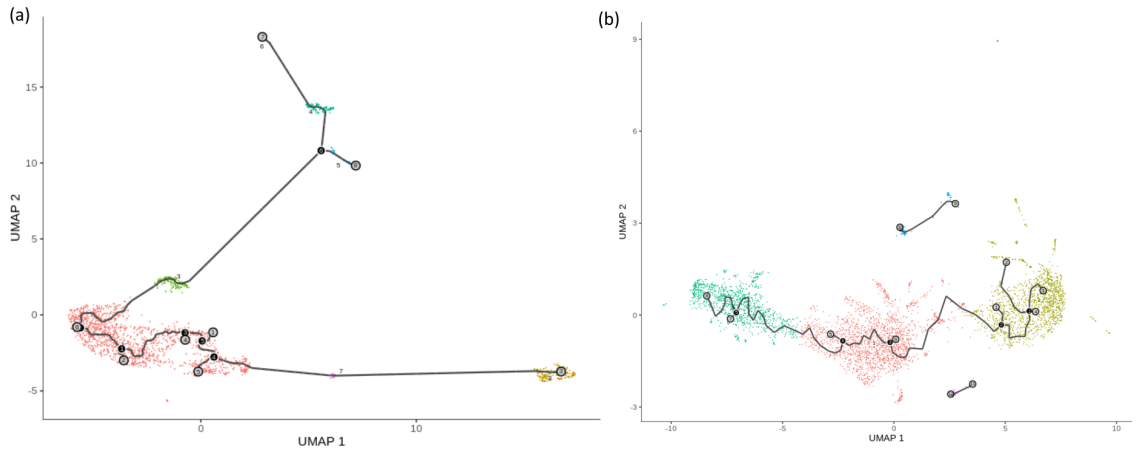
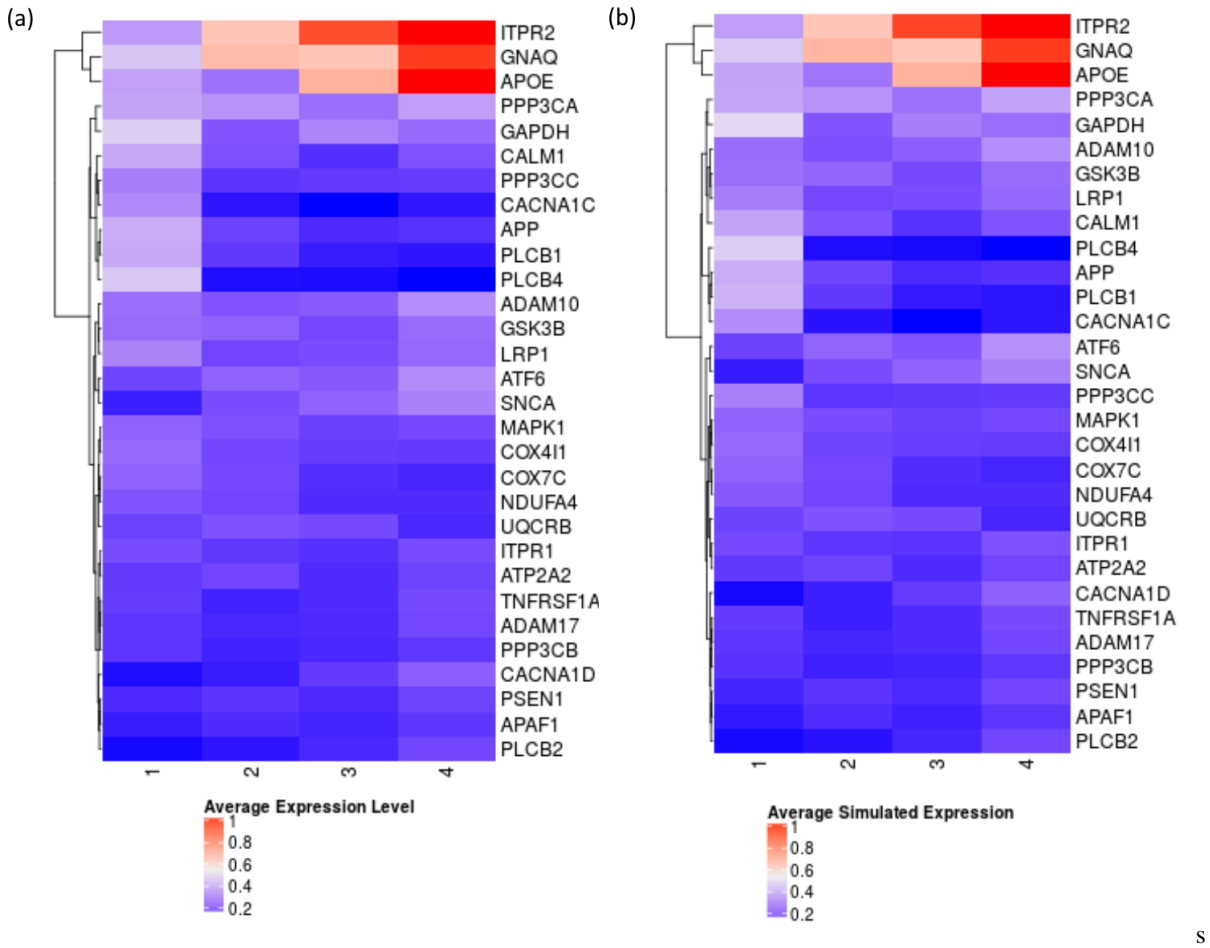


Figure S2: (a) Monocle clustering and trajectory of Alzheimer's microglial cells. (b) Monocle clustering and trajectory of Huntington's neuronal cells.



s

Figure S3: (a) Expression levels of 30 selected genes at each pseudotime point, averaged across the microglial cells which were assigned to that time point in the Alzheimer's analysis. (b) Average expression levels of 30 selected genes at each pseudotime point from 5,000 data points simulated using Poisson regression.

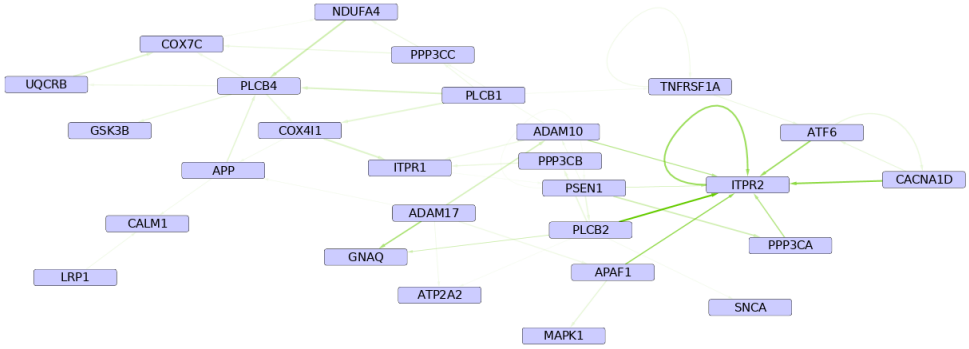


Figure S4: We used the disease score for each cell to construct a full pseudotemporal ordering, which we used as input to GRNVBEM [12] with the same 30 genes analyzed with our method. The resulting gene regulatory network was plotted in Cytoscape using the visual style provided in the author's GitHub repository

Method	Total Edges, 10 Iterations	Edges Duplicated At Least Once	Percent Duplicated
Fast Incremental Association	776	57	7.35
Incremental Association	770	62	8.05
Practical Constraint-Based	1198	121	10.10
Grow-Shrink	819	76	9.28
Interleaved Incremental Association	842	74	8.79
Incremental Association with FDR	6	0	0.00
Max-Min Parents and Children	1310	88	6.72
Hiton Parents and Children	2874	362	12.60
Hill Climbing	391	13	3.32
Tabu Search	407	16	3.93

Figure S5: Comparison of performance of several network learning methods tried. The result networks in this paper were learned using the Fast Incremental Association method.