Pathological cell state transitions in aging and disease in the human brain using single-cell RNA-seq

By Eduardo A. Maury

Introduction

Recent innovation in single-cell RNA-seq analyses has allowed for unprecedented insight into the developmental trajectory of multiple organs [1–3]. Cell-state transition prediction algorithms have been able to provide developmental trajectories of the prefrontal cortex at single cell resolution[4]. However, the potential of these novel methodologies has not been exploited in the study of the aging human brain and associated neurological disorders. The aging brain serves as an ideal application of cell state transition inference algorithms because of its limited cell proliferation capacity, which might be restricted to the dentate gyrus of the hippocampus if at all present [5,6]. Consequently, changes in the cell-state trajectory of neurons or glial cells in the adult brain can reflect physiological or pathological transitions secondary to aging or neurological disease, respectively.

Studying cell state transitions of the aging human brain can shed light on the pathophysiology of neurodegeneration. Alzheimer's disease (AD) is a progressive neurodegenerative condition resulting in dementia and memory loss. Pathologically it is characterized by accumulation of misfolded proteins tau and amyloid-beta [7–9] . Mouse models have allowed for a better understanding of the effect of these toxic proteins on neuronal populations, such as activation of apoptotic pathways [10–12]. However, how neurons go from a healthy state to a diseased state throughout the course of the disease remains unknown. In addition, mouse studies have shown the potential role of glial cell activation in AD [2,13]. However, the sequence of activation and cell state transitions of glial cells remains unknown in human brains. Histological examination of neurodegenerative processes such as AD, show that not all neurons or glial cells are involved by the toxic protein aggregates [8]. It follows that random sampling of single cells from a brain region could provide RNA-seq data to map out a continuum of pathological transition states.

Consequently, I propose to use cell-state trajectory mapping analysis to 1) identify the physiological cell transitions that occur during aging, and 2) determine the pathological deviations that occur in Alzheimer's disease across different cell types. The goal of this analysis is to discover key differentially expressed genes enriched in the transition to the pathological state, which might provide for new venues of therapeutic intervention.

In accordance with these aims, the analysis will have two stages. First, I propose to build a reference single-cell trajectory of aging brain cells using published algorithms that leverage graph abstractions to preserve global topology and allow for disconnected structures. This graph can be used to statistically determine control cell type marker genes. Then, using a similar approach I could re-cluster with brain cells from AD patients and statistically compare disease trajectories to control trajectories by using the marker genes previously discovered to guide cell identification. I could then further test for diverging disease genes by testing for differentially expressed genes across different diseased paths.

Aim1: Single-cell state trajectories in the aging brain

To interpret the cell state transitions that one might observe in disease, it is imperative to first understand the transitions in normal aging. Identification of marker genes, genes that are enriched at a branch point compared to others, are essential to characterize specific cell lineages and their developmental paths. These genes can be discovered statistically through single-cell trajectory construction. These marker genes can then be used as references for deviations in the disease state. Due to the fast pace of algorithmic development of single cell trajectory inference algorithm, there are robust and efficient approaches for single cell RNA-seq data [14].

1.1: Data generation via single cell RNA-seq of brain cells.

Previous single cell RNA-sequencing studies on human brain cells tend to be small, contain only a subpopulation of brain cells, and/or processed via different methodologies [4,15]. Consequently, it is necessary to generate new single cell RNA-seq data from human brains spanning several ages.

There are many parameters to consider for obtaining a representative sample that would capture the inherent heterogeneity from individual to individual. The first parameter to consider is the location of the brain to study. Focusing the analysis on the pre-frontal cortex, Broadmann Area 9, will provide a region without active neurogenesis and no AD pathology in cognitively normal individuals. An area of interest in the field of aging and AD is the hippocampus due to its role in memory formation. However, there is still debate on whether this area undergoes neurogenesis in adulthood [5,6]. Neurogenesis in adulthood might make it difficult to interpret single cell trajectories in that region since how neurogenesis in adulthood affects the transcriptome of neuronal populations remains unknown. To have stable trajectories, cells in this analysis should be in a relative quiescent stable state. In addition, the hippocampus is the first region of the brain to be affected in AD [8]. Consequently, even cognitively normal, aged individuals might have some tau pathology burden in this region, confounding our downstream analysis. Thus, using the prefrontal cortex is a better region for this study. The samples should come from fresh frozen brain tissue of individuals that died of causes unrelated to brain pathology such as accidents, non-neurological disease, or abortion for the fetal tissue. This will be accomplished with appropriate IRB approval and collaborations with neuropathologists.

The ages of the subjects should include key neurodevelopmental landmarks. Namely, this project aims to obtain brains at prenatal, childhood (ages 2-10), adolescence (ages 12-19), adulthood (ages 25-50), and aged (ages 80+) [16]. By the time the human brain reaches 25 years of age most of the brain is myelinated and cortical development is finalized. Consequently, it is expected that single cell trajectory reconstructions would fix on these stable adult/aged brain cell populations. Single cell and bulk RNA seq studies have indicated that some transcriptional changes occur in aging [2,13,17,18]. By including aged brains on this cohort, I will be able to capture normal variations within adult brain cell populations that might be undergoing physiological aging processes.

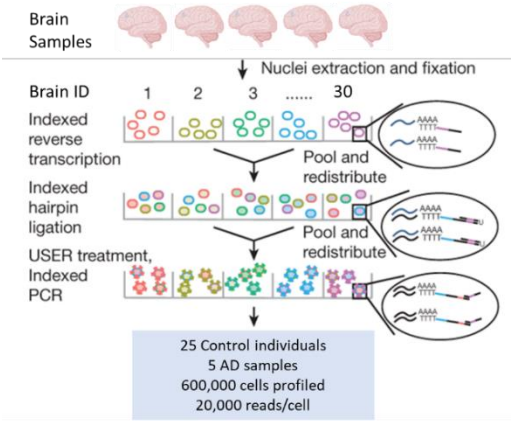

Figure 1 *Schematic of Sci-seq3 approach adapted from Cao et al. 2019*

While the ages above capture the natural aging variation, it is necessary to attempt to capture the person to person variability as much as possible. A recent study tried to assess the burden of somatic mutation through aging in single neurons [19]. In this study they sampled 2-6 subjects on each age category. With this sample size they were able to capture enough of the individual variability to observe trends with age. Since I aim to capture a wider group of cell types, setting a sample size of 5 individuals per group might be enough, for a total sample size of 25 subjects. To process the samples, I propose to use an approach similar to sci-RNA-seq3 (figure 1) which has been shown to allow for the profiling of 2 million cells in a single experiment through the use of combinatorial indexing [1]. In this experiment 61 mouse embryos staged between 9.5 and 13.5 days of gestation were used. They were able to keep information about the mouse of origin of each cell through the combinatorial indexing. I aim to obtain around 20,000 cells per individual, which would be in line with previous cell numbers in mice and human single cell trajectory analyses, yet it would provide unprecedented resolution in adult human cortex studies[1,4,13,20]. For the overall total of 500,000 cells, I propose to obtain ~20,000 reads per cell so allow for

detection of small transcription variations. To perform all these in a single experiment it would require the use of around 5 lanes in a Novaseq S4 flow cell, which produces around 2,250 million reads per lane. It would cost around $45,000-$50,000, which corresponds to around $10 per cell. The cost could be optimized by conducting pilot studies and assessing sample variability, to possibly reduce the total number of samples needed.

1.2: Single-cell trajectory analysis of aging cells using PAGA

The past few years has seen an explosion in the development of algorithms for single-cell trajectory inference from single-cell RNA-seq data[14,21,22]. A recent review compared the pros and cons of methods popular in the scientific community and designed recommendations for choosing an algorithm depending on the proposed plan of study, such as whether gaps in the trajectory were expected, and whether global topology should be preserved[14]. Single-cell RNA sequencing studies suffer from biases secondary to sampling. Differences in cell abundances might be secondary to which part of the tissue one is sequencing. As a result, it is expected that the inferred trajectories might have gaps. To allow for these discontinuities and yet robustly preserve global transcriptional relationships, I propose to use PArtition-based Graph Abstraction (PAGA), which provides a graph-like map of the arising data manifold, based on connectivity estimation of manifold partitions [22]. In brief, after standard preprocessing and quality filtering of reads, PAGA extracts highly variable genes for graph construction. The graph is built by constructing a symmetrized kNN-like graph using an approximate nearest neighbor search as described in UMAP, using Euclidean distance as the distance metric [23]. The graph is weighed using adaptive Gaussian kernels. One of the advantages of using PAGA is that it naturally allows for graph partitioning via a Louvain algorithm implementation, which can allow for data exploration a different resolution, from tissue, to cell types, to subtypes [1,22]. The portioning above can then be used to generate graphs using the test statistic quantifying the ratio of the number of inter-edges between clusters, normalized with the number of interedges expected under random assignments of edges [22]. Once a suitable graph is built, pseudotime estimation is done by adjusting for possible disconnected graphs as described in [24] and implemented in PAGA via random walk estimation. The PAGA graph can then be used to initialize established manifold learning and graph drawing algorithms such as UMAP for visualizations of single-cell embeddings. Notice that, pseudotime accuracy can be assessed since age information is readily available for each cell due to the combinatorial indexing from sci-RNA-seq.

Once the trajectory map is built, it is possible to discover marker genes of developmental and aging brain cell populations. Besides using already well-known cell type markers [1,4], I propose to implement a statistical framework similar to Chen et al [21] to discover diverging and transition genes. In this proposal, I define diverging genes as those that are expressed more at a particular branch, while transition genes are those that are enriched at particular timepoints along a branch.

To detect diverging genes, for each pair of diverging branches *Bi* and *Bj*, and for each gene *E* in those branches, the gene expression values can be normalized to range from 0 to 1. For a gene to be considered it needs to have a log2 fold change greater than a specific threshold, which Chen et al. chose to be 0.25 to enrich for highly variable genes. For each pair of gene expression E_i from B_i, and gene expression E_j from B_j, a Mann-Whitney U test can be performed to test hypotheses: *Ho:* $E_i=E_j$; *Ha:* $E_i \neq E_j$. Since in this analysis it is expected to have

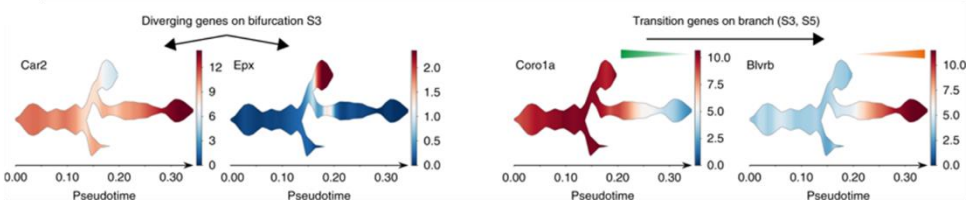

Figure 2*.Example of transition and divergent genes from Chen et al 2019*

large enough samples of cells across most branches, the U statistic of the Mann-Whitney test

can be approximated with a normal distribution by standardizing the U statistic to a Z score. This standardization will allow each U statistic to be comparable across different datasets. Fold change enrichment can be reported for the branches that have few cells (<20 cells per branch). Then genes with a z-score or fold change greater than a specific threshold are considered as diverging genes and enriched for a branch. The recommended threshold by Cao et al. is 2.

To detect transition genes, each gene expression for gene E for a branch Bi is scaled from 0 to 1. I can then calculate the log2 fold change in mean gene expressions of the first 20% and the last 80% of the cells based on the inferred pseudotime (and/or the known age categories), along Bi. If the difference is greater than a specified threshold (recommended to be 0.25), the Spearkman's rank correlation between inferred pseudotime and gene expression can be calculated for all the cells along the branch. Genes with Spearman's coefficients above a threshold (recommended 0.4) can be labeled as transition genes.

Overall, the identification of the diverging and transitions genes will serve as markers to label normal developing and aging populations. This analysis might provide novel marker genes of human brain cells, possibly illuminating new biology of aging and senescence. Paths characterized by these genes can then be compared for deviations along disease states as described below.

Aim 2: Single-cell trajectories in the diseased brain.

Alzheimer's disease is characterized by a highly stereotyped pathological progression, which makes it an ideal test case for trajectory analysis. AD pathology starts with the accumulation of tau in the enthorhinal cortex of the temporal lobe, and then progresses slowly to synapticaly connected regions in the limbic system, only reaching the prefrontal cortex in the later stages disease [8,9]. The pathological progression is classified by the Braak stage and CERAD score, which measure progression of tau pathology and amyloid beta distribution respectively. It has been shown that tau progression is the most correlated measure to clinical severity [9]. This slow progression results in the latest affected regions such as the prefrontal cortex to have sparse involvement, not all neuron or glial cells are affected by the pathology, as can be seen by histological evaluation. As a result, sampling from the prefrontal cortex of advanced AD cases (Braak stage 6) will result in a mixture of normal and diseased cells, which can then be separated by trajectory inference analysis.

2.1: Data generation via single cell RNA-seq

To minimize variability among samples and ensure generalization of the results, I propose to use a similar approach for collecting specimens from AD patients as described in Aim 1 section 1.1. Obtaining five samples from prefrontal cortex of pathologically proven AD patients will allow for assessment of the person to person variability between patients. However, it is not expected to have too much variability between patients since the disease follows a highly stereotyped progression which can be further controlled by procuring samples at the same Braak stage and CERAD score. Fresh frozen samples will be procured under appropriate IRB protocol from neuropathologist collaborators. In this study I propose to focus on late-onset AD subjects since sporadic AD is by far the most common form the disease, thus making tissue collection more feasible. To avoid batch effects, I propose to sequence the aged matched AD samples at the same time as the aging study samples. This modification will add one more lane to the study for a total two flow cells needed. In total I aim to sequence around 20,000 cells from each AD brain at 20,000 reads per cell. The NovaSeq has a two flow-cell capability, which will still allow for the joint sequencing of all the samples on the same experiment. The expected cost would be ~$55,000. Generation of this dataset would provide a unique resource to the scientific community, as such a large sequencing dataset run uniformly across disease and control samples has not been made publicly available to the best of my knowledge.

2.2: Single cell trajectory of Alzheimer's disease and comparison with control trajectories.

To build the single cell trajectory of AD, I propose to re-run the PAGA algorithm with the addition of the AD cells. By using this graph, we can easily visualize any possible deviations from normal aging cells by incorporating the sample-of-origin information provided by the combinatorial indexing. Benchmarking of PAGA for 1.3 million neuronal cells from mice of 10x genomics had a run time 90 seconds, for reference t-SNE takes about 10h and UMAP-only takes 191 minutes for the same dataset [22]. Consequently, re-running the single cell trajectory algorithm and data exploration can be performed efficiently with minimal computational costs. The clusters can then be labelled and inspected based on the gene markers discovered in the previous aim.

To test the hypothesis of whether the resulting aging graph, G*a, is different to the AD graph G*d, I propose to use the continuous measure of agreement used in Wolf et al to assess PAGA's robustness, which is illustrated in figure 3. In brief, this approach first computes the overlaps of the partitions labelled on each graph, generating an association matrix. This matrix can be normalized with respects to the groups in each graph. By taking the minimum of both normalizations, the minimal overlap is obtained for each group pair between graphs. In this way, each group in G*a can be associated with a group in G*d. Groups that do not have a corresponding association might indicate cell groups that are depleted or newly derived in AD.

Comparing the paths of the abstracted graphs G*a and G*d can reveal deviations in AD. For each shortest path between two leaf nodes in G*d, there is a shortest path between the associated nodes in G*a. The fraction of consistent steps between two paths can then be computed. To measure the agreement of the topologies between two abstracted graphs, we compute the fraction of agreeing steps and the fraction of agreeing paths over all combinations of leaf nodes in two given abstracted graphs. A similar approach was used in Wolf et al. Paths that do not have any corresponding agreement in G*a from G*d, can then be interrogated for diverging or transition genes as described in Aim 1 section 2.

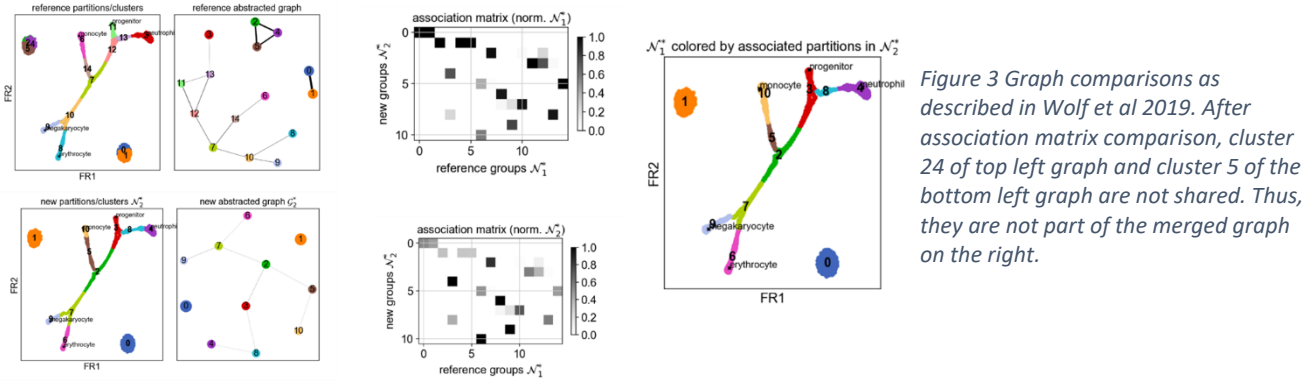

Figure 3 Graph comparisons as described in Wolf et al 2019. After association matrix comparison, cluster 24 of top left graph and cluster 5 of the bottom left graph are not shared. Thus, they are not part of the merged graph on the right.

It is possible that no differences exist between the single cell trajectories of AD and normal aging. This result would suggest that cell toxicity in AD might occur at the protein level at a time scale that does not perturb the transcriptome of cells. The cells might be suddenly overwhelmed and undergo cell death without time to invoke an appropriate transcriptional program response. However, previous transcriptomics studies of AD suggest that transcriptional changes can be caused by AD pathology [13,17,25]. Thus, I expect these changes to be reflected in the single cell RNA-seq trajectory reconstruction.

While the present study aims to uncover novel molecular understanding of aging and Alzheimer's disease, this approach could be broadly used in disease biology. The results of these analysis could serve as a framework to expand the field of single-cell state trajectory prediction to other areas of inquiry besides development. Using this type of analysis will provide a unique resolution to study how a normal cell transitions to a pathological state, which is fundamental to design appropriate therapies that delay or prevent disease onset.

Bibliography

1. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566,** 496–502 (2019).

2. Mathys, H. *et al.* Temporal Tracking of Microglia Activation in Neurodegeneration at Single-Cell Resolution. *Cell Rep.* **21,** 366–380 (2017).

3. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6,** (2017).

4. Zhong, S. *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555,** 524–528 (2018).

5. Moreno-Jiménez, E. P. *et al.* Adult hippocampal neurogenesis is abundant in neurologically healthy subjects and drops sharply in patients with Alzheimer's disease. *Nat. Med.* **25,** 554–560 (2019).

6. Sorrells, S. F. *et al.* Human hippocampal neurogenesis drops sharply in children to undetectable levels in adults. *Nature* **555,** 377–381 (2018).

7. Holtzman, D. M., Morris, J. C. & Goate, A. M. Alzheimer's disease: the challenge of the second century. *Sci. Transl. Med.* **3,** 77sr1 (2011).

8. Braak, H. & Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* **82,** 239–259 (1991).

9. Hyman, B. T., Van Hoesen, G. W., Damasio, A. R. & Barnes, C. L. Alzheimer's disease: cell-specific pathology isolates the hippocampal formation. *Science* **225,** 1168–70 (1984).

10. Pooler, A. M. *et al.* Amyloid accelerates tau propagation and toxicity in a model of early Alzheimer's disease. *Acta Neuropathol. Commun.* **3,** (2015).

11. Wegmann, S. *et al.* Removing endogenous tau does not prevent tau propagation yet reduces its neurotoxicity. *EMBO J.* **34,** (2015).

12. Wegmann, S. *et al.* Tau protein liquid-liquid phase separation can initiate tau aggregation. *EMBO J.* **37,** e98049 (2018).

13. Keren-Shaul, H. *et al.* A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell* **169,** 1276–1290.e17 (2017).

14. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* 1 (2019). doi:10.1038/s41587-019-0071-9

15. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 7285–90 (2015).

16. Chiurazzi, P. & Pirozzi, F. Advances in understanding – genetic basis of intellectual disability. *F1000Research* **5,** 599 (2016).

17. Guo, C. *et al.* Tau Activates Transposable Elements in Alzheimer's Disease. *Cell Rep.* **23,** (2018).

18. Twine, N. A., Janitz, K., Wilkins, M. R. & Janitz, M. Whole Transcriptome Sequencing Reveals Gene Expression and Splicing Differences in Brain Regions Affected by Alzheimer's Disease. *PLoS One* **6,** e16266 (2011).

19. Lodato, M. A. *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359,** 555–559 (2018).

20. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32,** 381–386 (2014).

21. Chen, H. *et al.* STREAM: Single-cell Trajectories Reconstruction, Exploration And Mapping of omics data. *bioRxiv* 302554 (2018). doi:10.1101/302554

22. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20,** 59 (2019).

23. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4314

24. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13,** 845–848 (2016).

25. Annese, A. *et al.* Whole transcriptome profiling of Late-Onset Alzheimer's Disease patients provides insights into the molecular changes involved in the disease. *Sci. Rep.* **8,** 4282 (2018).