scRNA-seq dropouts serve as a signal for tissue heterogeneity in autism spectrum disorder

by

Collin Spencer

Georgia Institute of Technology
Spring 2020

scRNA-seq dropouts serve as a signal for tissue heterogeneity in autism spectrum disorder

Approved by:

Dr. Greg Gibson, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Peng Qiu
School of Biomedical Engineering
Georgia Institute of Technology

Table of Contents

Abstract

Analysis of single-cell RNA-sequencing (scRNA-seq) data is plagued by dropouts, zero counts for mRNA transcripts due to low mRNA in individual cells and inefficient mRNA capture. Dropouts are traditionally treated as an error to be corrected through normalization while performing unsupervised clustering of single cells based on highly expressed, variable transcripts. A novel algorithm, co-occurrence clustering, treats dropouts as a signal and binarizes scRNA-seq data for cell clustering, producing the same clusters as Seurat.

Previous application of Seurat to single nuclear RNA-sequencing (snRNA-seq) data taken from the prefrontal cortex (PFC) and anterior cingulate cortex (ACC) of patients with autism spectrum disorder (ASD) found no difference in clusters between brain regions. This seems at odds with literature suggesting tissue-specific emergence of co-expression networks and regional specialization in the brain. We applied co-occurrence clustering to ASD samples to parse interregional heterogeneity between the PFC and ACC and identify novel cell clusters.

Introduction

Autism spectrum disorder (ASD) affects 1.9% of children in the United States. Complex gene-environment interactions give rise to ASD, and recent studies suggest that at least 50% of the risk of developing ASD is attributable to genetic variation [1]. Despite this, ASD and other psychiatric disorders are highly heterogeneous, which makes identification of important genetic risk factors difficult. Historically, approaches involving genome-wide association studies (GWAS) to identify common genetic variants in the form of single-nucleotide polymorphisms (SNPs) have been successful [2]. However, common genetic variants individually explain just a small amount of the variance in liability, and cumulatively genome-wide significant SNPS account for only 1-2% of observed variance [3]. As such, no singular genetic variant can be attributed to ASD. Nevertheless, from the array of genes identified with relevance to ASD, gene networks have been constructed that show convergence on neuronal development, modulation, and intracellular transcriptional mechanisms.

Gene expression profiling by RNA-seq can be used to identify cell types associated with ASD, offering an additional layer of granularity. Indeed, evidence of convergent dysregulation from cell types has been revealed in the first of its kind study using single-nucleus RNA sequencing (snRNA-seq) [4]. Postmortem samples of ASD and control patients were taken from the Anterior Cingulate Cortex (ACC) and Prefrontal Cortex (PFC) and processed for nuclei isolation and snRNA-seq using the 10x Genomics platform. Unbiased clustering of the ACC and PFC together and separately resulted in the identification of the same cell types. This appears to be at odds with literature suggesting tissue-specific co-expression networks emerge spatiotemporally from different cell types [5]. Recent advances in neural circuits using the research domain criteria (RDoC) shows that the PFC and ACC serve together and separately in mood and anxiety disorders [6]. Significant genetic overlap with ASD and hierarchical transcriptomic specialization across brain regions suggest that some distinction should be apparent between the PFC and ACC [7].

Potential explanations for why this was not captured in the original study include sample size and dropout events. Dropout events are a limitation of single-cell RNA-seq studies, arising from low mRNA in individual cells and inefficient mRNA capture [8]. In existing scRNA-seq methods, feature selection is typically performed only once through principal component analysis of highly variable genes. In this pipeline, dropouts are filtered out and the remaining transcript abundance counts are normalized such that only highly expressed variable genes are considered. A new co-occurrence clustering algorithm developed by Dr. Peng Qiu at Georgia Tech treats these dropouts as a useful signal instead of noise to identify additional cell types. Furthermore, feature selection is re-visited for each iteration of clustering to characterize the heterogeneity of cells under consideration. An initial assessment of this approach shows that co-occurrence clustering produces more clusters than previous analyses of the same datasets. This suggests that different bioinformatic strategies may be optimal for analyzing cell populations with varying levels of heterogeneity, such as those found in the PFC and ACC.

To further understand region-specific cellular heterogeneity in ASD, this study will examine how dysregulation of the ACC and PFC contribute to ASD through the application of co-occurence clustering to scRNA-seq samples. It is predicted that this method will identify additional cell types implicated in tissue-specific gene networks through subsequent differential gene expression analyses of the ACC and PFC. This novel approach seeks to further characterize the heterogeneous nature of ASD, identifying novel dysregulated pathways for therapeutic intervention and diagnosis.

1. Yoo H. (2015). Genetics of Autism Spectrum Disorder: Current Status and Possible Clinical Applications. *Experimental neurobiology*, *24*(4), 257–272. doi:10.5607/en.2015.24.4.257
2. Losh, M., Sullivan, P. F., Trembath, D., & Piven, J. (2008). Current developments in the genetics of autism: from phenome to genome. *Journal of neuropathology and experimental neurology*, *67*(9), 829–837. doi:10.1097/NEN.0b013e318184482d
3. Velmeshev D, et. al. (2019). Single-cell genomics identifies cell type-specific molecular changes in autism. *Science*. 685-689.

4. Le, B. D. and Stein, J. L. (2019), Mapping causal pathways from genetics to neuropsychiatric disorders using genome‐wide imaging genetics: Current status and future directions. Psychiatry Clin. Neurosci., 73: 357-369. doi:10.1111/pcn.12839
5. Williams, Leanne M (2017). Precision psychiatry: a neural circuit taxonomy for depression and anxiety. *The Lancet Psychiatry*, Volume 3, Issue 5, 472 - 480
6. Burt J, et al. (2018). Hierarchy of transcriptomic specialization across human cortex captured by myelin map topography. *Nature Neuroscience*.
7. Qiu P (2019). Embracing the dropouts in single-cell RNA-seq data. bioRxiv.

Literature Review

Clinical heterogeneity is a hallmark of autism spectrum disorder (ASD) as reflected in the current Diagnostic and Statistical Manual for Mental Disorders (DSM-5). Likewise, recent research initiatives to detect the biological basis of psychiatric disorders have revealed significant genetic heterogeneity that underpins phenotypic expression[1]. However, the path from gene to behavior is complex and arduous. Network-based approaches seek to integrate the common and rare genetic variants involved in autism and identify cellular pathways [2]. Recent advances in single-cell RNA sequencing (scRNA-seq) have introduced a new level of granularity in this approach. Unbiased clustering of expression leads to the identification of cell types involved in disease pathogenesis. Subsequent differential gene expression analysis of cell types and gene ontology analysis constructs gene networks that provide new insights into ASD through the identification of affected cellular pathways. Despite these advances, dropout reads from scRNA-seq data acquisition and increasing evidence of region-specific transcriptomic profiles in the brain pose barriers, and present opportunities for further research [3,4]. To address this, a novel co-occurrence clustering algorithm has been developed that uses dropout reads as a signal [5]. This approach may detect intra-region heterogeneity with greater precision than traditional methods, potentially yielding new insight into therapeutic targets for ASD.

An overarching goal of current research is to define ASD by its genetic components. At least 50% of the risk of developing ASD is attributable to genetic variation [7]. Hundreds of genes are likely related to autism. Unfortunately, genome-wide association studies (GWAS) have failed to identify single variants at genome-wide significance [6]. It is only within the past several years that sample sizes have grown large enough to detect the minuscule effect sizes of pathogenic variants on the disease itself. One explanation for why GWAS has failed to identify additional variants lies in the focus of GWAS on susceptibility loci, which ignores other variants that are not causative alone [8]. Connections between genes and behaviors are further complicated by how etiology is shared across multiple complex disorders. Indeed, results from the Psychiatric Genomics Consortium show a significant overlap between variants associated with schizophrenia and autism [8]. Furthermore, epigenetic factors can have a profound impact on the transcriptome of an organism. Epigenetic factors may be causative in of themselves but may also influence the expression of risk genes to modulate disease state [7,8]. In the journey from gene to behavior, gene-gene and gene-environment interactions result in hierarchical disease manifestations across the genome, transcriptome, proteome, etc. It is thus imperative to explore each of these -omic analyses both independently and jointly, in order to uncover how ASD behavior emerges from lower biological levels.

Co-occurence clustering of dropout reads may yield an additional level of precision by detecting layers of complexity to neural circuit composition that are not apparent from studies of only highly abundant genes. Large-scale neural circuits serve as endophenotypes and theoretically lie closer to the underlying biology of ASD. The central idea is that neural circuits compute behavioral actions and that this computation is heavily influenced by the physical architecture of neuronal connectivity encoded by genetic variants [9]. One example of a large-scale neural circuit is negative valence, which interprets the intrinsic averseness of environmental stimuli [9]. Negative valence systems involve connectivity and activation between the anterior cingulate cortex, insula, and amygdala. Because neural circuits define physical regions of the brain, studying which variants are associated with altered architecture should yield dysfunctional pathways. However, several well-powered genetic studies of psychiatric disorders have failed to demonstrate more than a handful of correlations between genetic variants and structural imaging traits derived from fMRI scans [10]. A likely explanation for this is that pathogenic variants do not significantly alter gross brain structure, but act at the cellular or subcellular level. One line of evidence to support this hypothesis comes from the study of transcriptomic specialization across the human cortex. Hierarchical transcript gradients define an axis shared by transcriptomic and anatomical architecture of the cortex [11]. Transcriptomic diversity influences microscale properties that contribute to macroscale function. With the advent of single-cell RNA sequencing, it is now possible to characterize individual cell types within micro and macro neural circuits that contribute to ASD.

One of the advantages of co-occurence clustering using single-cell RNA is the level of precision it yields compared to traditional bulk-brain sequencing. Single-cell RNA sequencing has introduced an unprecedented level of granularity in identifying how neuropsychiatric conditions affect individual cell types [12]. Single-nucleus RNA sequencing analysis identified that synaptic signaling of upper-layer excitatory neurons and microglia are preferentially affected in autism [13]. However, few differentially expressed genes were observed compared to bulk-brains sequencing and even fewer between regions sampled (prefrontal cortex and anterior cingulate cortex). One potential explanation is due to the infamous issue of dropouts. Limitations of scRNA-seq include low capture efficiency and sequence coverage that results in a higher level of noise than bulk RNA-seq data [14]. Dropouts occur when no transcript is captured. This issue is particularly pernicious in regions of high cell-to-cell heterogeneity, where dropout events increase. Dropouts are typically handled through normalization when focusing on highly variable genes (e.g. Seurat) or used to perform feature selection through methods like M3Drop [14-16]. A new co-occurrence clustering algorithm developed by Dr. Qiu uses dropouts as a useful signal similar to M3Drop [17,18]. Unlike M3Drop, however, feature selection is re-visited for each iteration of clustering to produce more clusters than traditional methods like Seurat. Because of the method's ability to handle varying levels of heterogeneity, co-occurrence clustering may be particularly well-suited to the transcriptomic diversity and specialization of neural circuits. As such, the application of co-occurrence clustering to characterize intraregional heterogeneity in ASD samples should detect novel pathways that are otherwise obscured by focusing on highly expressed, variable genes.

In sum, genetic variation affects multiple levels of biology, at different developmental periods, within certain cell types, altering the physical structure of neural circuits, affecting how circuits perform computations, and producing canonical behavioral symptoms of ASD [9,10]. By researching how ASD affects individual cell types within neural circuit regions, it may be possible to identify new pathways for disease treatment and identification. Crucial to this endeavor will be how to handle dropouts that influence data quality and, ultimately, the accuracy of implicated pathways. Co-occurrence clustering and similar methods that treat dropouts as a useful signal may complement or outperform existing strategies that focus on highly expressed variable genes. The unprecedented precision these methods offer into how ASD emerges at the cellular level will help to redefine ASD by its biological components and identify new therapeutic targets.

References

1. Demkow, U, and T Wolańczyk. “Genetic tests in major psychiatric disorders-integrating molecular medicine with clinical psychiatry-why is it so difficult?.” *Translational psychiatry* vol. 7,6 e1151. 13 Jun. 2017, doi:10.1038/tp.2017.106
2. Grimes, T., Potter, S. S., & Datta, S. (2019). Integrating gene regulatory pathways into differential network analysis of gene expression data. *Scientific reports*, *9*(1), 5479. doi:10.1038/s41598-019-41918-3
3. Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, *50*(8), 96. doi:10.1038/s12276-018-0071-8
4. Sunkin, Susan M et al. “Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system.” *Nucleic acids research* vol. 41,Database issue (2013): D996-D1008. doi:10.1093/nar/gks1042
5. (2018, November 17). Embracing the dropouts in single-cell RNA-seq data | bioRxiv. Retrieved October 18, 2019, from https://www.biorxiv.org/content/10.1101/468025v2
6. Grove, J., Ripke, S., Als, T. D., Mattheisen, M., Walters, R. K., Won, H., … Børglum, A. D. (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nature genetics*, *51*(3), 431–444. doi:10.1038/s41588-019-0344-8
7. Yoo H. (2015). Genetics of Autism Spectrum Disorder: Current Status and Possible Clinical Applications. *Experimental neurobiology*, *24*(4), 257–272. doi:10.5607/en.2015.24.4.257

8. Rylaarsdam, L., & Guemez-Gamboa, A. (2019). Genetic Causes and Modifiers of Autism Spectrum Disorder. *Frontiers in cellular neuroscience*, *13*, 385. doi:10.3389/fncel.2019.00385

9. Williams L. M. (2016). Precision psychiatry: a neural circuit taxonomy for depression and anxiety. *The lancet. Psychiatry*, *3*(5), 472–480. doi:10.1016/S2215-0366(15)00579-9

10. Le, B. D. and Stein, J. L. (2019), Mapping causal pathways from genetics to neuropsychiatric disorders using genome‐wide imaging genetics: Current status and future directions. Psychiatry Clin. Neurosci., 73: 357-369. doi:10.1111/pcn.12839

11. Burt, J. B., Demirtaş, M., Eckner, W. J., Navejar, N. M., Ji, J. L., Martin, W. J., … Murray, J. D. (2018). Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography. *Nature neuroscience*, *21*(9), 1251–1259. doi:10.1038/s41593-018-0195-0

12. Bryois, Julien & Skene, Nathan & Hansen, Thomas & Kogelman, Lisette & Watson, Hunna & Brueggeman, Leo & Breen, Gerome & Bulik, Cynthia & Arenas, Ernest & Hjerling Leffler, Jens & Sullivan, Patrick. (2019). Genetic Identification of Cell Types Underlying Brain Complex Traits Yields Novel Insights Into the Etiology of Parkinson's Disease: Supplementary Tables. 10.1101/528463.

13. Single-cell genomics identifies cell type–specific molecular changes in autism

14. By Dmitry Velmeshev, Lucas Schirmer, Diane Jung, Maximilian Haeussler, Yonatan Perez, Simone Mayer, Aparna Bhaduri, Nitasha Goyal, David H. Rowitch, Arnold R. Kriegstein

    Science17 May 2019 : 685-689

15. Chen, G., Ning, B., & Shi, T. (2019). Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Frontiers in genetics*, *10*, 317. doi:10.3389/fgene.2019.00317

16. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, *33*(5), 495–502. doi:10.1038/nbt.3192

Methods

snRNA-seq data from ASD patients was obtained from a previous study conducted by Velmeshev et al and available at the SRA accession PRJNA434002. Information about nuclei extraction and quality control are available in Velmeshev's supplementary materials. Samples were split into groups based on the cortical region they were obtained from (PFC or ACC) and phenotype (ASD or control).

Library demultiplexing, fastq file generation, read alignments, and unique molecular identifiers (UMI) quantification was achieved using CellRanger software v 1.3.1 and default parameters. Samples were aligned against the pre-mRNA reference file (ENSEMBL GRCh38) to capture introns. After processing by CellRanger, we performed two separate methods of dimensionality reduction, clustering, and visualization through a MATLAB-based implementation.

Samples from the PFC and ACC for ASD and control patients are run through the MATLAB-based co-occurrence clustering algorithm to produce cell clustering results. A full description of co-occurrence clustering is provided by Qiu on biorxiv (https://www.biorxiv.org/content/10.1101/468025v2.full.pdf+html). For the purposes of this paper, we will provide a brief summary of the methodology.

First, the algorithm binarizes the scRNA-seq count matrix and evaluates a statistical measure for co-occurrence between each pair of genes through the Jaccard index. The graph is then partitioned into gene clusters using the Louvain algorithm, leaving gene clusters with high co-occurrence that represent pathways for major groups of cell types in heterogeneous populations. Percentages of detected genes in each cell are used to build a cell-cell graph. Community detection is then applied to further partition the graph into smaller clusters in an iterative manner. Cluster validity is assessed by mean difference and mean ratios. Similar clusters are merged while heterogeneous ones continue to be partitioned into subclusters.

After cell clustering and visualization with co-occurrence clustering, the same samples are clustered using a MATLAB-based Seurat implementation. A confusion heatmap compares the cell clusters generated by co-occurrence clustering and Seurat to demonstrate the difference in the algorithms' ability to partition heterogeneity in cortical samples.

Results

One sample clustering

Application of co-occurrence clustering to sn-RNAseq data gathered from the PFC of one ASD patient produced 34 cell clusters (Fig. 1), and Seurat identified 15 (Fig. 2).

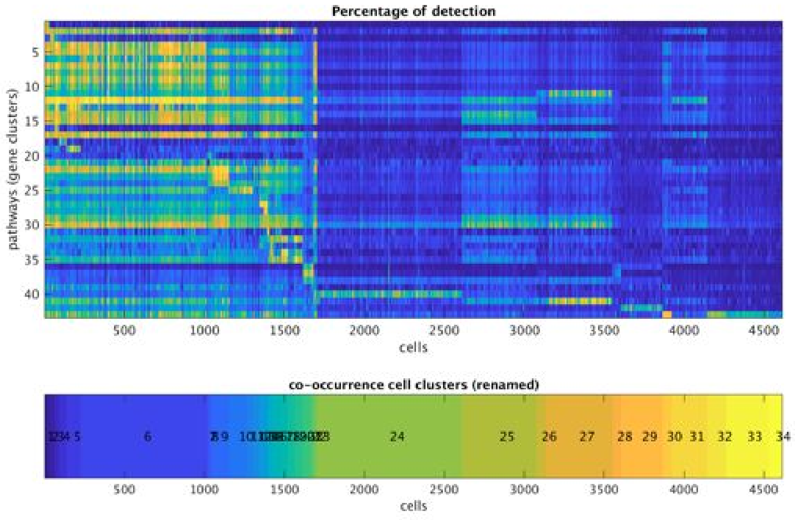

Figure 1: Co-occurrence clustering of one ASD PFC sample

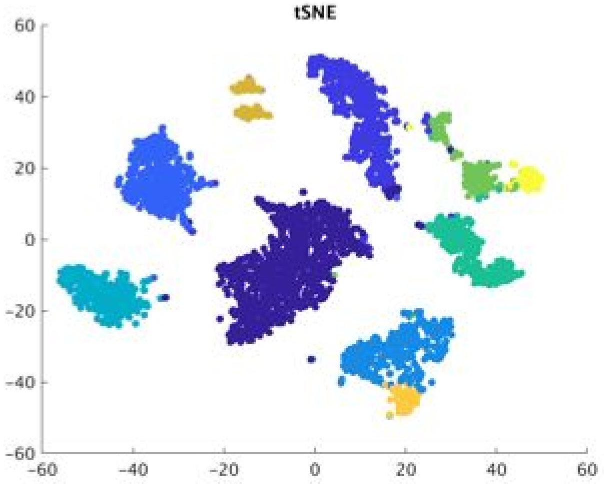

Figure 2: Seurat clustering and tSNE representation of one ASD PFC sample

Aggregate Sample Clustering

Co-occurrence clustering was performed on aggregate samples from the ACC and PFC of patients with ASD. We found 119 robust cell clusters in the PFC and 98 clusters in the ACC. Furthermore, 142 gene clusters were found in the PFC compared to 125 clusters in the ACC.

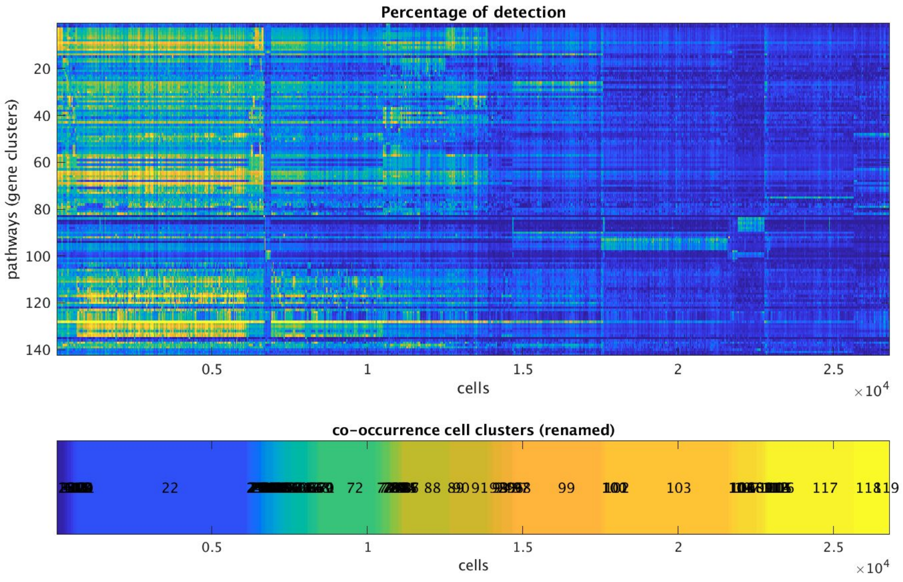

Figure 3: Co-occurence clustering of ASD PFC samples

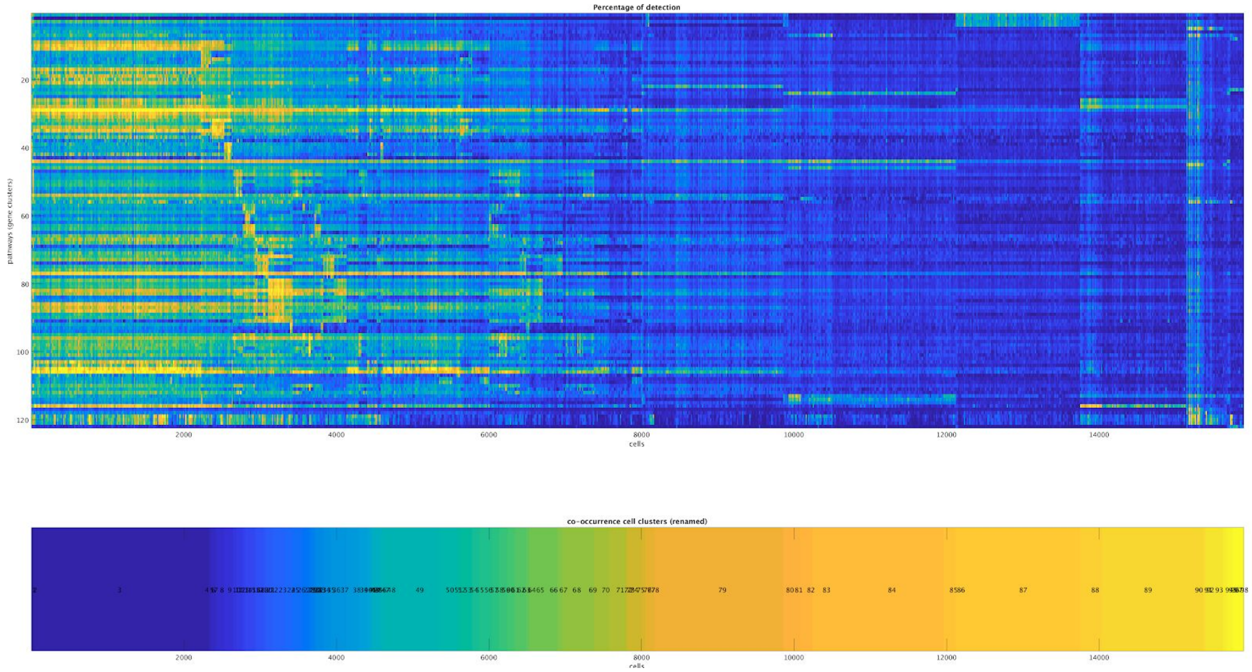

Figure 4: Co-occurence clustering of ASD ACC samples

Seurat Clustering

Seurat with tSNE for dimensionality reduction was used to cluster cells within the PFC (Figure 5) and ACC (Figure 6), identifying 20 clusters for both regions. A MATLAB-based implementation of Seurat, as part of the co-occurence clustering pipeline, was used to perform dimensionality reduction and generate respective tSNE plots.

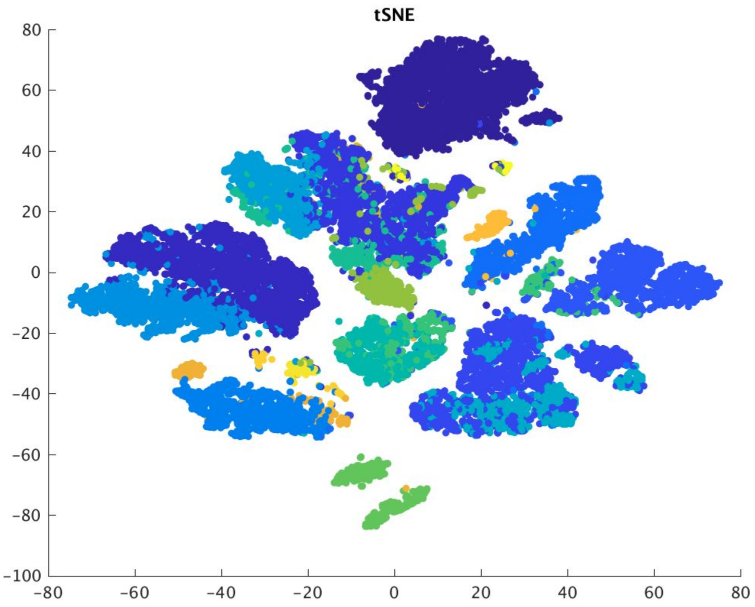

Figure 5: PFC Seurat clusters

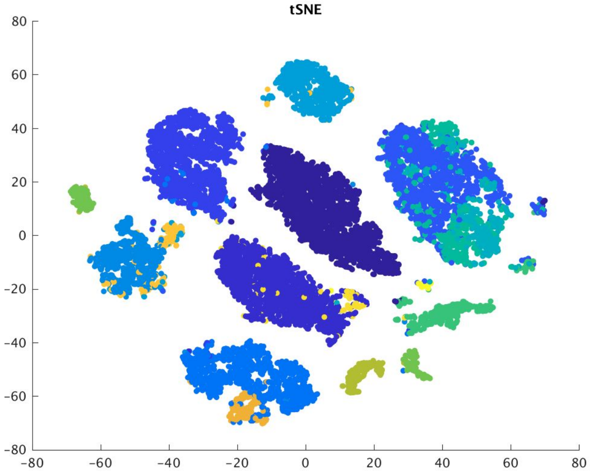

Figure 6: ACC Seurat clusters

Discussion

Our results demonstrate the utility of co-occurrence clustering for parsing interregional heterogeneity in the brain. Within a single sample of the PFC from one ASD patient, 34 cell clusters were identified by co-occurrence clustering (Fig. 1) compared to 15 by Seurat (Fig. 2). The high ratio of co-occurrence to Seurat clusters observed in a single sample generalizes to aggregate samples of the PFC and ACC. Seurat only identified 20 cell clusters for both regions (Fig. 5 and 6), whereas co-occurrence clustering produced 119 clusters in the PFC (Fig. 3) and 98 clusters in the ACC (Fig. 4). The identification of the same number of cell clusters by Seurat is consistent with prior analysis by Velmeshev et al [1]. The 5x difference in cluster count between Seurat and co-occurrence clustering confirms our initial hypothesis that co-occurrence, owing to its use of dropouts and iterative feature selection, would produce more clusters than Seurat. It is worth noting, however, that this is the first application of co-occurrence clustering to sn-RNAseq data. Therefore, it is possible that the high cluster ratio is, in part, a technical error for not adjusting parameters of co-occurrence clustering to the new data type and the pre-mRNA reference.

Co-occurrence clustering not only identified more clusters than Seurat, but produced a marked difference in cluster count between brain regions. Our observed discrepancy between cell clusters found in the PFC and ACC (119, 98 respectively) matches the disparity in gross brain volume between the two regions [2]. The larger volume of the PFC presents the opportunity for greater intraregional specialization, and thus more cell types and subtypes, compared to the ACC. Furthermore, additional studies have highlighted the region-dependent emergence and specialization of gene co-expression networks in the brain [3, 4]. The difference in cluster count between the ACC and PFC from co-occurrence clustering supports the existence of region-specific dysregulation in gene co-expression networks in the context of ASD.

In sum, we have demonstrated the utility of co-occurrence clustering for snRNA-seq data, identified novel cell clusters implicated in ASD, and reported the first-known delineation of interregional heterogeneity between PFC and ACC clusters from our data set. Our results show that dropouts, typically viewed as a source of technical error, can be used as a valuable signal to analyze highly heterogeneous regions of the brain. Further analysis of identified cell clusters may yield novel dysfunctional pathways unique to our sampled brain regions and improve our understanding of the biological emergence of ASD.

References

1. Velmeshev D, et. al. (2019). Single-cell genomics identifies cell type-specific molecular changes in autism. *Science*. 685-689.
2. Kiho Im, Jong-Min Lee, Oliver Lyttelton, Sun Hyung Kim, Alan C. Evans, Sun I. Kim .(2008). Brain Size and Cortical Structure in the Adult Human Brain. *Cerebral Cortex*. 2181–2191.
3. Burt J, et al. (2018). Hierarchy of transcriptomic specialization across human cortex captured by myelin map topography. *Nature Neuroscience*.
4. Williams L. M. (2016). Precision psychiatry: a neural circuit taxonomy for depression and anxiety. *The lancet. Psychiatry*, *3*(5), 472–480. doi:10.1016/S2215-0366(15)00579-9