Comparison of Methods for Differential Gene Analysis in Autism Using snRNA Data

THORDAR HAN[1], Cuining Liu[2], Katherine Eyring[3], Kevin Abuhanna[2], Yi Zhang[2], Brie Wamsley[3], Daniel H. Geschwind[2,3], Chongyuan Luo[2]

1 BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
2 Department of Human Genetics, David Geffen School of Medicine, UCLA
3 Department of Neurology, David Geffen School of Medicine, UCLA

ABSTRACT

One debate within the single-nucleus RNA sequencing (snRNA-seq) field involves whether to sequence many cells superficially or fewer cells deeply. To investigate this issue, we analyzed two single-nucleus human cortex datasets to detect cell-type-specific differentially expressed genes (DEGs) between autism spectrum disorder (ASD) and control groups. Dataset A profiled 83,958 nuclei (median 2,157 counts/nucleus) using 10X Genomics (Velmeshev et al., 2019). Dataset B profiled 18,666 nuclei (median 96,660 counts/nucleus) using snmCT-seq (Luo/Geschwind Lab). We investigated best practices for integrating these datasets and detecting DEGs. We observed that pseudobulk and cell-level DEG-identification methods have correlated effect estimates, but result in different sets of significant DEGs. Both datasets implicated multiple cell types in ASD with a surprising non-neuronal signature; however, Dataset B yielded a greater number of results, including SFARI-validated autism-related genes that were absent in Dataset A, such as SHANK3, CAMK2A, and UBE3A.

BACKGROUND

- **Single-nucleus RNA sequencing (snRNA-seq)** offers higher resolution transcriptome analysis than conventional bulk tissue approaches.
- **Data Collection: 10X Genomics versus snmCT-seq**
  - 10X Genomics uses droplet-based technology.
  - snmCT-seq is plate-based, capturing multiomic data but with lower throughput with respect to number of cells.
- **DEG Analysis: Pseudobulk versus cell-level modeling**
  - Pseudobulk aggregates RNA counts from multiple cells of the same type.
  - Cell-level modelling preserves individual cell heterogeneity.
    - While the original Velmeshev et al. paper used MAST, we used a comparable method called Dream.
- ASD research has Implicated genes like MEF2C, SHANK3, CAMK2A, and UBE3A that influence synaptic function and neural plasticity.
- However, non-neuronal cells, including glial and immune cells, are becoming increasingly associated with ASD pathology.

snmCT-seq

METHODS

Donor Demographics

Dataset A
(10X, Velmeshev)

		ASD	Control
Sample Size		12	12
Cells per Donor	Mean	3500.3	3496.2
Age	Mean (SD)	14.5 (4.9)	13.7 (5.6)
PMI	Mean (SD)	21.8 (10.8)	16.1 (5.0)
Region	ACC	7 (58.3)	8 (66.7)
	PFC	5 (41.7)	4 (33.3)

Dataset B
(snmCT, Luo/Geschwind)

		ASD	Control
Sample Size		15	13
Cells per Donor	Mean	618.8	721.8
Age	Mean (SD)	21.5 (13.2)	29.5 (17.1)
PMI	Mean (SD)	27.9 (15.8)	25.5 (11.7)
Region	PFC	15 (100.0)	13 (100.0)

Cell Counts Per Cluster

Workflow for Attempted Integration and DEG Analysis

Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., Bhaduri, A., Goyal, N., Rowitch, D. H., & Kriegstein, A. R. (2019). Single-cell genomics identifies cell type-specific molecular changes in autism. *Science (New York, N.Y.)*, *364*(6441), 685–689. https://doi.org/10.1126/science.aav8130

RESULTS

Dataset B Has More Features, Counts, and Higher Variance than Dataset A

Dataset A (VLMC)   **Dataset B (VLMC)**

Various Attempts at Dataset Integration Were Unsuccessful

Downsampled number of cells in Dataset A | Downsampled read counts in Dataset B | Integration with Harmony | Integration with RPCA

Final Clustering with Azimuth

Given poor integration, we used Azimuth human cortex reference data for cluster annotation.

Pseudobulk and Cell-Level Modelling are Well-Correlated

Dataset A

Pearson r̄ = 0.602

Dataset B

r̄ = 0.600

Effect Size Correlation Between Datasets was Higher in Cell-Level Modelling

Pseudobulk

r̄ = 0.038

Cell-Level Modelling

r̄ = 0.233

Correlation between datasets in cell-level modelling seems to be higher for clusters with more cells.

Dataset B Generally Identified More Significant DEGs

- Number of significant genes with total number of genes tested.
- In pseudobulk, Dataset B identified a greater proportion of significant DEGs* (p < 0.05) in all clusters.
- The difference is less pronounced for cell-level modelling, but Dataset A still has a lower absolute number and mean proportion of significant DEGs than Dataset B.

*Most clusters had no significant results when considering FDR adjusted p-values < 0.1

- Number of SFARI genes found significant with total number of SFARI genes included in analysis.
- In pseudobulk, Dataset B found a greater proportion of SFARI genes to be significant across all clusters.
- This trend was not as strong in cell-level analysis, but Dataset B still found a greater absolute number of SFARI genes significant.

Significant Autism-Related SFARI Genes* Only Found in Dataset B

Pseudobulk

CAMK2A (Exc L5 ET)
logFC = 6.12e-2
p-value = 0.798

logFC = -3.54
p-value = 1.04e-3

SHANK3 (Exc L6 CT)
logFC = -0.213
p-value = 0.490

logFC = -1.59
p-value = 4.99e-2

UBE3A (Exc L6 IT Car3)
logFC = 1.92e-3
p-value = 0.990

logFC = -2.21
p-value = 1.29e-2

Cell-Level Modelling

NFIB (Endo)
logFC = 4.48e-2
p-value = 0.771

logFC = 0.428
p-value = 3.42e-2

VEZF1 (Oligodendrocytes)
logFC = 6.97e-3
p-value = 0.718

logFC = 0.264
p-value = 2.39e-3

MKX (IN-SST)
logFC = 5.82e-2
p-value = 0.211

logFC = 0.356
p-value = 3.23e-2

- Of the 428 SFARI genes found significant in Dataset B but not Dataset A with pseudobulk:
  - 11 were filtered out for low number of reads.
  - 23 were filtered out for not being expressed in 50% of participants.
  - 226 were filtered out due to low variance.
- With cell-level modelling, 61 out of 76 SFARI genes found significant only in Dataset B were filtered out of Dataset A due to low variance, including RAC1, MEF2C, and ANKRD11.

*SFARI genes for this analysis were defined as either Category 1 or Syndromic SFARI genes.

CONCLUSIONS

- Integration with cell atlases like Azimuth is a useful method with the current limitations of available dataset-integration tools.
- Having deeper coverage of cells allows for the capture of validated DEGs that would be missed with shallow-level sequencing.
- Although pseudobulk is the default method for DEG analysis, cell-level modelling may be a more powerful option when the data consists of more cells.
- Gene expression differences related to autism were found in non-neuronal cells such as astrocytes and VLMC that warrant further investigation.