Towards better understanding of developmental disorders from integration of spatial single-cell transcriptomics and epigenomics

Guojie Zhong [*1] **Jiayao Wang** [*1 2] **Siyu He** [*3] **Xi Fu** [*2]

Abstract

The recent emerging techniques of single cell spatial RNA seq makes it possible to profile the transcriptomics data at single cell resolution without loss of the spatial information. However, it is still a challenge to measure epigenomics profiles at spatial levels. In this project, we developed an autoencoder based multi-omics integration method and applied it on spatial mouse fetal brain data to reconstruct the spatial epigenomics profiles. We compared our method with LIGER and showed its better performance on a public dataset measured by latent mixing metrics. We further developed a CNN model to predict autism risk genes based on the spatial RNA seq data. Our model is able to prioritize autism risk genes from whole genome level. Code of our project can be found at https://github.com/explorerwjy/ML_genomics.git

1. Introduction

Recent advances in single cell sequencing technologies are able to profile genetic, epigenomics and transcriptomic data at single cell resolution within the whole tissue or organs, providing the ability of identifying cell heterogeneity and dynamical developments. The emerging techniques of spatial transcriptomics complement the spatial information of the tissue, making it possible to study the cell-cell interaction in a systematic way. However, current methodologies haven't been able to measure epigenomics profiling at spatial levels, which is important for a comprehensive understanding of organogenesis and pathogenesis of human genetic disorders. Multimodal spatial techniques that simultaneously measure the spatial transcriptome and epigenomics in the same samples have not yet been developed, without which the accuracy of downstream analysis could be heavily affected by the batch and sample variations. Thus, there is a highly unmet need for developing an integrated method for simultaneously analyzing transcriptomic and epigenomic data in the perspective of both spatial and single cell level.

The overarching aim of this project is to develop new computational methods to integrate the spatial scRNA-seq and non-spatial scATAC-seq data and leverage such information to improve power of risk gene discovery in disease genetics studies.

2. Methods

2.1. Data sets and preprocessing

In this project, we integrated four datasets: scATAC-seq(Preissl et al., 2018), scRNA-seq(Cao et al., 2019), spatial RNA-seq (Liu et al., 2020) and exon sequencing (Feliciano et al., 2018). We mapped those scRNA-seq, spatial RNA-seq and scATAC-seq to the spatial coordinates of the brain regions and identified some spatial patterns may associated to Autism spectrum disorders (ASD).

The spatial RNA-seq dataset(Liu et al., 2020) used deterministic barcoding in tissue for spatial omics sequencing (DBiT-seq) for co-mapping of mRNAs and proteins in a formaldehyde-fixed tissue slide via next-generation sequencing. Gene expression profiles in 10 μm pixels conform to the clusters of single-cell transcriptomes, allowing for rapid identification of cell types and spatial distributions. DBiT-seq was conducted in 10 μm and 25 μm pixel size to analyze the whole embryo and brain region of mouse embryo (E10–E12), respectively. Data was downloaded from GEO with the accession number GSE137986.

The scATAC-seq dataset(Preissl et al., 2018) applied a combinatorial indexing assay to profile genome-wide chromatin accessibility in ∼100,000 single cells from 13 adult mouse tissues. They identify 85 distinct patterns of chromatin accessibility, most of which can be assigned to cell types,

[*]Equal contribution [1]Department of Systems Biology, Columbia University, Columbia University, New York, USA [2]Department of Biomedical Informatics, Columbia University, New York, USA [3]Department of Biomedical Engineering, Columbia University, New York, USA. Correspondence to: Guojie Zhong <gz2294@cumc.columbia.edu>, Jiayao Wang <jw3514@cumc.columbia.edu>, Siyu He <sh3846@columbia.edu>, Xi Fu <xf2217@cumc.columbia.edu>.

The 2021 ICML Workshop on Computational Biology.

and $\sim$ 400,000 differentially accessible elements. Data was downloaded from GEO with the accession number GSE111586. Peak signals were summarized into gene level (gencode vM25), by mapping peak genomic locations to gene's promoter regions (3k bp upstream of transcription start site) plus gene body(Welch et al., 2019).

The scRNA-seq dataset(Cao et al., 2019) profiled the transcriptomes of around 2 million cells derived from 61 embryos staged between 9.5 and 13.5 days of gestation in a single experiment. We downloaded the data from GEO with the accession number GSE119945, keeping 14 brain-related cell types in 10.5 day of gestation as annotated in the study.

The exon sequencing dataset(Feliciano et al., 2018) we are using is the largest exome sequencing study of autism spectrum disorder (ASD) to date (n = 35,584 total samples, 11,986 with ASD). They used an enhanced analytical framework to integrate *de novo* and case-control rare variation and identify 102 risk genes at a false discovery rate of 0.1 or less. Here we used their high confident genes as our candidate genes for ASD.

2.2. An autoencoder based method for integration of multi-omics datasets

We developed an autoencoder model for integration of multi-omics datasets. For each cell i in each batch k, the input will be the observed gene counts X_{ik}. The shared encoder NN_E will encode it to a d-dimensional latent space representation, Z_{ik}, which could be interpreted as the expression level of several "meta genes" of that cell. Different batches will share an encoder. For each batch k, the decoder part will consist of a shared decoder, NN_D, as well as a dataset specific NN_{B_k}, which will decode the latent space representation to two reconstructed output, W_{i_k} and V_{i_k}, correspondingly. W_{i_k} represents the shared characteristics through datasets while V_{i_k} represents the dataset specific characteristics, in other words, batch effect. The loss function is defined as:

$$\sum_i \sum_k \|V_{i_k} + W_{i_k} - X_{i_k}\|_2 + \lambda \|V_{i_k}\|_2 \qquad (1)$$

Under this loss function the encoder and decoder will be trained to minimize the dataset specific characteristics while keeping as much useful information as possible in the latent representation for reconstruction. The dimension of Z_{i_k}, d and the loss term λ, are the hyperparameters. In practice we use $d = 20$ and $lambda = 5$. For all the neural networks, we used one layer of fully-connected network followed by a ReLU activation layer. We implemented our model in pytorch and trained 20 epochs on PBMC(Zheng et al., 2017) data, 5 epochs on mouse fetal brain data. The latent representation of cells were used to compare with LIGER and perform downstream analysis.

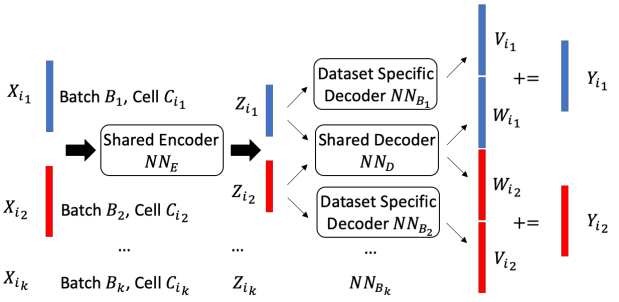


Figure 1. Autoencoder model

2.3. Latent mixing metric

The latent mixing metric was adopted from (Gayoso et al., 2021), which measures how well the latent cell representations are mixed between batches relative to the global frequency of batches. The entropy is calculated as:

$$KL(p^{(n)}||q) = \sum_{i=1}^{B} p_i^{(n)} log \frac{p_i^{(n)}}{q_i} \qquad (2)$$

where $p_i^{(n)}$denoted the frequency of cell type i in the 100 nearest neighbors of cell n. q_i denoted the global frequency of cell type i. We ran 50 times to get the average entropy, and in each run, we randomly selected 100 cells as the objects.

2.4. Cluster mixing metric

The cluster mixing metric measures how well the cells clustering is consistent with the biological cell type annotations. We first performed a graph-based clustering using Phenograph (Levine et al., 2015) on the latent representation of cells. The predicted cell type for each cluster was then defined as the most frequent cell types in each cluster. To evaluate the performance, We calculated the confusion matrix to compare the predicted cell type versus the original biological annotated cell types.

2.5. KNN to identify and validate cell type

We applied a KNN method to identify the spatial scRNA cell type from the annotated non-spatial scRNA-seq dataset. For each grid in spatial scRNA-seq, we chose the nearest k scRNA cells in the latent-cell matrix and used their majority annotated cell types as the prediction. We will compare this annotation with the brain anatomical annotations as an evaluation of the performance of LIGER and our Autoencoder model.

2.6. Phenograph to cluster cells and aligning the cell type

We utilized the Phenograph(Levine et al., 2015) method to cluster the cells of multi-omics data based on their shared cell embedding matrix. Phenograph also used KNN and Jaccard graphs with the Minkowski metric. We annotated the cell type of scATAC and spatial RNA based on the pre-annotated scRNA by determining the majority cell types in each cluster.

2.7. CNN to predict autism genes with spatial expression data

In order to investigate spatial patterns of autism risk genes, we used a convolutional neural network model to predict disease risk. We collected 88 well known autism risk genes from the SFARI (Simons Foundation Autism Research Initiative) Gene database(Banerjee-Basu & Packer, 2010). which are genes strongly implicated in autism based on expert curation from the literature. For negative genes, we collected 977 genes with at least 1 *de novo* LGD (likely-gene disrupting) or missense variant in unaffected siblings and no mutations in probands, from an exome-sequencing study(Iossifov et al., 2014). Since we are using mouse developing spatial RNA-seq data, we map the genes to mouse homologs(Hayamizu et al., 2005)(Smith & Eppig, 2009), which leaves 85 and 853 autism positive/negative genes respectively. Spatial RNA-seq data on the E10 mouse brain of each gene was transformed to 50 x 50 x 1 tensors, each pixel is one spatial spot that DBiT-seq has measure expression on. We normalized the UMI counts to (0,1) scale, since we would like to capture spatial patterns rather than the expression level of those genes. In order to access the model performance, we split the data as 70% training and 30% testing set, and for data in training set, we produce pseudo-training data by adding small amount of UMI to random spatial spot of real data, to increase sample size, to the model would be robust to noise and prevent overfitting. For convolutional neural net models, we used 2 convolutional blocks followed by 2 linear fully connected layers. Each convolutional block consists of a 2D-convolutional layer, a batch normalization layer, one ReLU and one Max Pooling layer. Before each fully connected layer we also used a dropout layer with drop out rate 0.2. Model architecture shown in Fig 5A. Considering the fact of very small sample size, we used a simple model architecture with heavy regularization.

3. Experimental Results

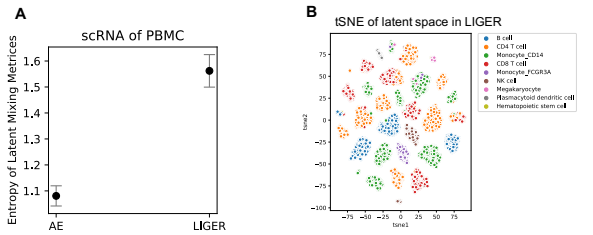
3.1. Autoencoder model performed better than LIGER in latent entropy

To evaluate the performance of autoencoder and LIGER, we selected two types of metrics: latent mixing metric and cluster mixing metric. We first tested the latent mixing metric in PBMC(Zheng et al., 2017) datasets, which the cell type annotation has been well studied and evaluated. Figure 2A demonstrated the latent entropy of autoencoder was lower than LIGER, which indicated the relationship between cell type and latent was less disordered in autoencoder, and latent space of autoencoder presented the cell types better than LIGER. Then we performed cluster mixing metric by measuring the accuracy of cell type annotation. Figure 2B and Figure 2C presented the tSNE map of latent space of LIGER and autoencoder colored by real cell types, and Figure 2D and Figure 2E showed the tSNE map of latent space of LIGER and autoencoder colored by clusters identified by Phenograph. We plotted the confusion matrix (Figure 2F and Figure 2G) to evaluate the accuracy of predicted cell types. Figure 2H, Figure 2I showed the classification reports as well as the accuracy and F1 scores. Though the overall accuracy of the autoencoder is worse than LIGER, it still got a 0.73 accuracy and F1 score 0.86 and 0.92 at B cell and Monocyte CD4 cells.

3.2. Spatial mapping of single cell data

We first performed the latent mixing metric on the single cell RNA expression of an E10 mouse embryo. The results demonstrated the lower entropy of AE method than LIGER (Figure 3A). analyzed the spatial transcriptome of an E10 mouse embryo from the DBiT-seq with a resolution of 25um. Figure 3B showed the bright field of the mouse brain region, including nostrils, telencephalon, mesencephalon, rhombencephalon. We first preprocessed the spatial RNA data and plotted the UMI counts on the same map (Figure 3C). The sequencing depth is acceptable and comparable with scRNA. Figure 3D-F exhibited several examples of latent distribution of autoencoders on spatial transcriptome. Some patterns were indicated in the images. Figure 3G and Figure 3H plotted the tSNE of the single cell colored by the clusters determined by phenograph. By the method of KNN as described above, we obtained the maps of cell type in the mouse brain region (Figure 3I). Based on the results, excitatory neurons have not developed much yet in the E10 embryo. Then we also plotted the tSNE map of scRNA, scATAC, and spatial RNA colored by the cell types and datasets type (Figure 3J). The result indicated the spatial RNA and scATAC conformed well with scRNA. Meanwhile, we also used Phenograph to cluster the cells to identifying the cell types based on the pre-annotated cell types in scRNA data. Both methods showed similar neuron

type maps among brain regions. Surprisingly, we identified a bunch of sensory neurons around the region of the trigeminal ganglia, which is considered as the sensory ganglion of the trigeminal nerve that occupies a cavity in the dura mater.



H

	precision	recall	f1-score	support
B cell	0.99	1.00	0.99	2371
CD4 T cell	0.95	0.93	0.94	4450
CD8 T cell	0.90	0.94	0.92	3142
Hematopoietic stem cell	0.00	0.00	0.00	24
Megakaryocyte	0.66	0.58	0.62	106
Monocyte_CD14	0.97	0.98	0.98	4090
Monocyte_FCGR3A	0.96	0.88	0.92	561
NK cell	0.96	0.96	0.96	593
Plasmacytoid dendritic cell	1.00	0.99	0.99	139
accuracy			0.95	15476
macro avg	0.82	0.81	0.81	15476
weighted avg	0.95	0.95	0.95	15476

I

	precision	recall	f1-score	support
B cell	0.93	0.80	0.86	2371
CD4 T cell	0.61	0.90	0.73	4450
CD8 T cell	0.58	0.30	0.40	3142
Hematopoietic stem cell	0.00	0.00	0.00	24
Megakaryocyte	0.00	0.00	0.00	106
Monocyte_CD14	0.87	0.97	0.92	4090
Monocyte_FCGR3A	0.85	0.33	0.48	561
NK cell	0.53	0.41	0.46	593
Plasmacytoid dendritic cell	0.00	0.00	0.00	139
accuracy			0.73	15476
macro avg	0.49	0.41	0.43	15476
weighted avg	0.72	0.73	0.70	15476

Figure 2. Evaluation of autoencoder (AE and LIGER with PBMC data. A) Entropy of AE and LIGER B) tSNE of latent space in LIGER, colored by cell type. C) tSNE of latent space in AE, colored by cell type. D) tSNE of latent space in LIGER, colored by Phenograph cluster. E) tSNE of latent space in AE, colored by Phenograph cluster. F) Confusion matrix of predicted types by LIGER and true types. G) Confusion matrix of predicted types by AE and true types. H) Classification report of LIGER. I) Classification report of AE.

Another spatial RNA data that zooms in the eye region confirmed the same identification of the sensory ganglion clusters, with a higher resolution (Figure 3J). Meanwhile, we noticed a large proportion of oligodendrocyte progenitors near the eye regions, corresponding to the area of optic stalk. Radial glia and neural tubes have identified as the largest percentages of neural cells. Postmitotic premature neurons were identified near the hinder brain and cerebral cortex, which correspond to the specification of excitatory neurons. Figure 3K showed the cell annotation by autoencoder. The result was different with what LIGER acquired, but autoencoder identified a large quantity of neural tubes, which played an important role at the brian development. Figure 3L and Figure 3M demonstrated the cluster map of the factor matrix of cells in data with AE and LIGER respectively. It indicated the shared latent features of some cells, corresponding to the cell type specific features, and also indicated the shared features among scRNA, spatial RNA and scATAC. Then we mapped the scATAC data by finding the closest nearest neighbor in each spatial spot of spatial transcriptome, and mapping latent space of ATAC on to spatial RNA space. Superisely, we figured out the latent space of scATAC still exhibited spatial resolution and informations, which further indicated the usefulness of studying spatial features of ATAC sequencing, such as exploring the risk genes at spatial resolution. Overall, the results suggested that our autoencoder performed pretty well in finding latent features by integration of multi-omics data, and able to recapitulate some specific patterns of the developmental processes.

3.3. Autism genes and prediction with spatial expression patterns

In order to visualize expression patterns of autism candidate genes, we select candidate genes from the SFARI dataset(Banerjee-Basu & Packer, 2010). We compute the expression specificity of each spatial location as a Z score of UMI compared to all other spatial locations, for the same gene, And expression specificity of the candidate gene set as the average of Z-scores of all candidate genes. As shown in Figure 4, We notice mean ASD gene expression specificity (Figure 4A) has a high correlation with axonogenesis and hindbrain neuron development, as well as telencephalon development and regionalization(Liu et al., 2020). We also notice that ASD genes have distingic spatial expression patterns (Figure 4B) . Some genes like Adnp or Dscam, are specifically expressed in certain locations, while other genes like Chd8 and Dyrk1a, show a ubiquitous expression across the fetal brain.

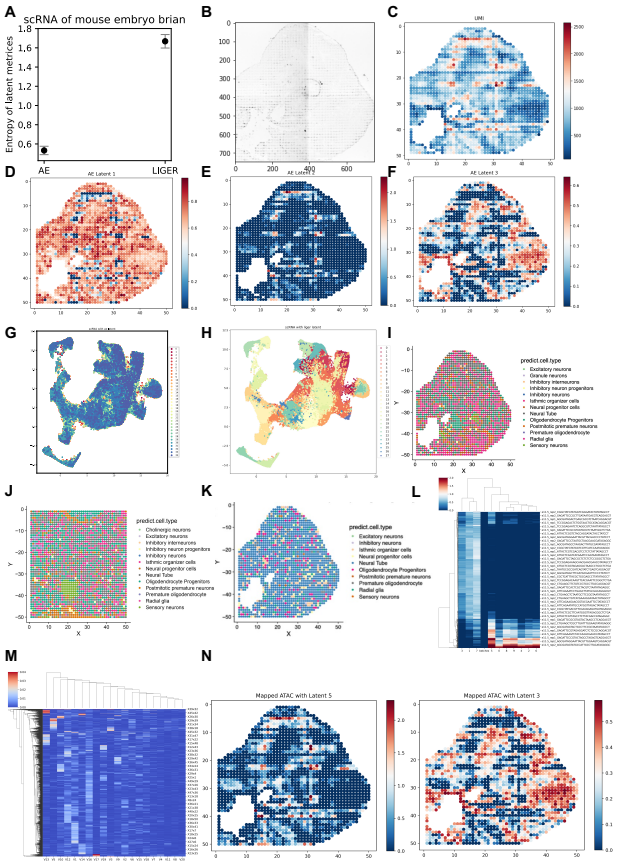


Figure 3. Spatial RNA and ATAC prediction in the E10 mouse brain region. A) Entropy of AE and LIGER with E10 mouse brain data. B) bright field of E10 mouse brian region. C) UMI distribution of DBiT-seq data. D-F) Distribution of latent variables at dimension 1/2/3 by autoencoder. G) tSNE plot colored by cell clusters (by Phenograph) on the LIGER latent. H) tSNE plot colored by cell clusters (by Phenograph) on the Autoencoder latent. I) Predicted cell type on E10 mouse brain by LIGER. J) Predicted cell type on E10 mouse eye region by LIGER. K) Predicted cell type on E10 mouse brain by Autoencoder. L) Cluster map of AE latent. M) Cluster map of LIGER latent. N) Mapped spatial ATAC latent distribution.

With evidence that autism genes do have distinct spatial expression patterns, we used a simple CNN to learn autism spatial expression features and predict autism genes based on that. Figure 5A shows the architecture of CNN model. Considering the small number of well known autism genes we can use as positive training data, we only used a 2-layer CNN with a heavy regularied model. Despite that, the model is still overfitting on training data. As Figure 5B shows, the model achieves AUC = 0.944 on training data but only AUC = 0.738 on the testing set. In order to evaluate our scores's ability to distinguish autism genes, we used a largest published dataset on autism trio exome sequencing(Rodgers et al., 2016). We calculate *de novo* enrichment (observed number of mutations divided by expected number of mutations) of groups of 1000 genes, genes were sorted by our prediction score. We took out genes that appear in training data to prevent overestimating our score performance. From Figure 5C and D, we can see genes with larger prediction scores have a larger burden of both LGD and damaging missense mutations, which means our scores are able to pick out autism risk genes. For example, the top 1000 genes with the highest prediction score have an LGD burden of 2.16, much greater than the whole genome level of 1.33. Interestingly, the top gene group didn't have the highest damaging missense score (defined by MPC score > 1(Ke et al., 2017)), the 3rd highest gene group has the largest burden of 1.55, compared to genome wide level of 1.11, which means autism genes may have different contribution from LGD and damaging missense mutations.

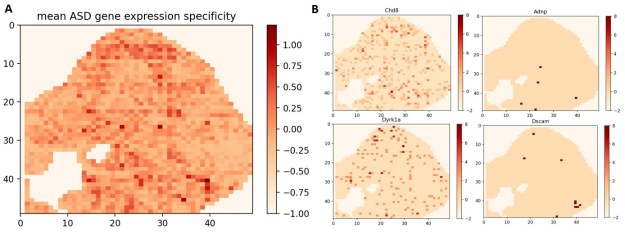


Figure 4. Spatial expression of Autism genes in the fetal brain. (A): Expression specificity of all ASD candidate genes and (B): Expression specificity of selected ASD candidate genes

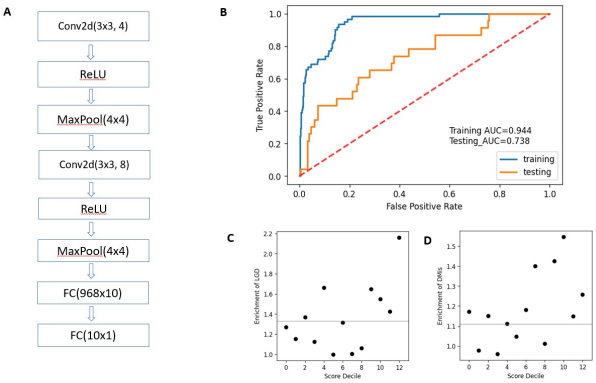


Figure 5. Autism risk predicted by CNN. (A):CNN architecture (B): ROC curve of prediction on held out testing set. (C) : LGD enrichment of genes sorted by prediction score. (D): Damaging missense mutations (DMis) enrichment of genes sorted by prediction score. From right to left are groups of genes with high to low prediction scores. Horizontal lines represent genome-wide burden of each mutation class (with training autism genes removed).

4. Discussion

In this project, we developed a new multi-omics integration method and compared it with LIGER. Our model showed higher latent mixing metrics while slightly worse clustering mixing metrics compared to LIGER, which could be due to the Phenograph clustering algorithm. We have to notice although autoencoder is a powerful tool to learn latent representations of the data, there is still space to for improvement to get a better performance in more applications.

Our CNN model is, to our best knowledge, the first model that utilizes 2D spatial expression patterns of genes in predicting disease risk genes. The biggest challenge of these types of tasks is lack of training data. CNN is a powerful machine learning tool for capturing features in 2D space but also requires a large amount of training data. There are 20,000 protein coding genes in total, among which dozens have been proved associated with Autism, out model can still work in prioritizing the rest genes with limited data points. Our results suggest that genes associated with autism do have spatial patterns in early embryo development, which can be used to further understand the disease mechanism.

References

Banerjee-Basu, S. and Packer, A. SFARI Gene: an evolving database for the autism research community. *Disease Models & Mechanisms*, 3(3-4):133–135, April 2010. ISSN 1754-8411. doi: 10.1242/dmm.005439.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C., and Shendure, J. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, February 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-0969-x. URL https://www.nature.com/articles/s41586-019-0969-x. Number: 7745 Publisher: Nature Publishing Group.

Feliciano, P., Daniels, A. M., Green Snyder, L., Beaumont, A., Camba, A., Esler, A., Gulsrud, A. G., Mason, A., Gutierrez, A., Nicholson, A., Paolicelli, A. M., McKenzie, A. P., Rachubinski, A. L., Stephens, A. N., Simon, A. R., Stedman, A., Shocklee, A. D., Swanson, A., Finucane, B., Hilscher, B. A., Hauf, B., O'Roak, B. J., McKenna, B., Robertson, B. E., Rodriguez, B., Vernoia, B. M., Van Metre, B., Bradley, C., Cohen, C., Erickson, C. A., Harkins, C., Hayes, C., Lord, C., Martin, C. L., Ortiz, C., Ochoa-Lubinoff, C., Peura, C., Rice, C. E., Rosenberg, C. R., Smith, C. J., Thomas, C., Taylor, C. M., White, L. C., Walston, C. H., Amaral, D. G., Coury, D. L., Sarver, D. E., Istephanous, D., Li, D., Nugyen, D. C., Fox, E. A., Butter, E. M., Berry-Kravis, E., Courchesne, E., Fombonne, E. J., Hofammann, E., Lamarche, E., Wodka, E. L., Matthews, E. T., O'Connor, E., Palen, E., Miller, F., Dichter, G. S., Marzano, G., Stein, G., Hutter, H., Kaplan, H. E., Li, H., Lechniak, H., Schneider, H. L., Zaydens, H., Arriaga, I., Gerdts, J. A., Cubells, J. F., Cordova, J. M., Gunderson, J., Lillard, J., Manoharan, J., McCracken, J. T., Michaelson, J. J., Neely, J., Orobio, J., Pandey, J., Piven, J., Scherr, J., Sutcliffe, J. S., Tjernagel, J., Wallace, J., Callahan, K., Dent, K., Schweers, K. A., Hamer, K. E., Law, J. K., Lowe, K., O'Brien, K., Smith, K., Pawlowski, K. G., Pierce, K. L., Roeder, K., Abbeduto, L. J., Berry, L. N., Cartner, L. A., Coppola, L. A., Carpenter, L., Cordeiro, L., DeMarco, L., Grosvenor, L. P., Higgins, L., Huang-Storms, L. Y., Hosmer-Quint, L., Herbert, L. M., Kasparson, L., Prock, L. M., Pacheco, L. D., Raymond, L., Simon, L., Soorya, L. V., Wasserburg, L., Lazar, M., Alessandri, M., Brown, M., Currin, M. H., Gwynette, M. F., Heyman, M., Hale, M. N., Jones, M., Jordy, M., Morrier, M. J., Sahin, M., Siegel, M. S., Verdi, M., Parlade, M. V., Yinger, M., Bardett, N., Hanna, N., Harris, N., Pottschmidt, N., Russo-Ponsaran, N., Takahashi, N., Ousley, O. Y., Juarez, A. P., Manning, P., Annett, R. D., Bernier, R. A., Clark, R. D., Landa, R. J., Goin-Kochel, R. P., Remington, R., Schultz, R. T., Brewster, S. J., Booker, S., Carpenter, S., Eldred, S., Francis, S., Friedman, S. L., Horner, S., Hepburn, S., Jacob, S., Kanne, S., Lee, S. J., Mastel, S. A., Plate, S., Qiu, S., Sandhu, S., Thompson, S., White, S., Myers, V. J., Singh, V., Yang, W. S., Warren, Z., Amatya, A., Ace, A. J., Chatha, A. S., Lash, A. E., Negron, B., Rigby, C., Ridenour, C., Stock, C. M., Schmidt, D., Fisk, I., Acampado, J., Nestle, J. L., Nestle, J. A., Layman, K., Butler, M. E., Kent, M., Mallardi, M. D., Carriero, N., Lawson, N., Volfovsky, N., Edgar, R., Marini, R., Rana, R., Ganesan, S., Shah, S., Ramsey, T., Chin, W., Jensen, W., Krentz, A. D., Gruber, A. J., Sabo, A., Salomatov, A., Eng, C., Muzny, D., Astrovskaya, I., Gibbs, R. A., Han, X., Shen, Y., Reichardt, L. F., and Chung, W. K. SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. *Neuron*, 97(3):488–493, February 2018. ISSN 0896-6273. doi: 10.1016/j.neuron.2018.01.015. URL https://www.sciencedirect.com/science/article/pii/S0896627318300187.

Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., and Yosef, N. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3):272–282, March 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01050-x. URL https://www.nature.com/articles/s41592-020-01050-x. Number: 3 Publisher: Nature Publishing Group.

Hayamizu, T. F., Mangan, M., Corradi, J. P., Kadin, J. A., and Ringwald, M. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. *Genome*

Biology, 6(3):R29, February 2005. ISSN 1474-760X. doi: 10.1186/gb-2005-6-3-r29. URL https://doi.org/10.1186/gb-2005-6-3-r29.

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman, H. A., Witherspoon, K. T., Vives, L., Patterson, K. E., Smith, J. D., Paeper, B., Nickerson, D. A., Dea, J., Dong, S., Gonzalez, L. E., Mandell, J. D., Mane, S. M., Murtha, M. T., Sullivan, C. A., Walker, M. F., Waqar, Z., Wei, L., Willsey, A. J., Yamrom, B., Lee, Y.-h., Grabowska, E., Dalkic, E., Wang, Z., Marks, S., Andrews, P., Leotta, A., Kendall, J., Hakker, I., Rosenbaum, J., Ma, B., Rodgers, L., Troge, J., Narzisi, G., Yoon, S., Schatz, M. C., Ye, K., McCombie, W. R., Shendure, J., Eichler, E. E., State, M. W., and Wigler, M. The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, 515 (7526):216–221, November 2014. ISSN 1476-4687. doi: 10.1038/nature13908. URL https://www.nature.com/articles/nature13908. Number: 7526 Publisher: Nature Publishing Group.

Ke, S., Ja, K., Kj, K., Ah, O.-L., E, P.-H., Dg, M., Bm, N., and Mj, D. Regional missense constraint improves variant deleteriousness prediction. June 2017. doi: 10.1101/148353. URL https://europepmc.org/article/PPR/PPR28042.

Levine, J., Simonds, E., Bendall, S., Davis, K., Amir, E.-a., Tadmor, M., Litvin, O., Fienberg, H., Jager, A., Zunder, E., Finck, R., Gedman, A., Radtke, I., Downing, J., Pe'er, D., and Nolan, G. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162(1):184–197, July 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.05.047. URL https://www.sciencedirect.com/science/article/pii/S0092867415006376.

Liu, Y., Yang, M., Deng, Y., Su, G., Enninful, A., Guo, C. C., Tebaldi, T., Zhang, D., Kim, D., Bai, Z., Norris, E., Pan, A., Li, J., Xiao, Y., Halene, S., and Fan, R. High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. *Cell*, 183(6):1665–1681.e18, December 2020. ISSN 00928674. doi: 10.1016/j.cell.2020.10.026. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867420313908.

Preissl, S., Fang, R., Huang, H., Zhao, Y., Raviram, R., Gorkin, D. U., Zhang, Y., Sos, B. C., Afzal, V., Dickel, D. E., Kuan, S., Visel, A., Pennacchio, L. A., Zhang, K., and Ren, B. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nature Neuroscience*, 21(3):432–439, March 2018. ISSN 1546-1726. doi: 10.1038/s41593-018-0079-3. URL https://www.nature.com/articles/s41593-018-0079-3. Number: 3 Publisher: Nature Publishing Group.

Rodgers, J., Wigham, S., McConachie, H., Freeston, M., Honey, E., and Parr, J. R. Development of the anxiety scale for children with autism spectrum disorder (ASC-ASD). *Autism Research*, 9(11):1205–1215, 2016. ISSN 1939-3806. doi: https://doi.org/10.1002/aur.1603. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aur.1603. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aur.1603.

Smith, C. L. and Eppig, J. T. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *WIREs Systems Biology and Medicine*, 1(3):390–399, 2009. ISSN 1939-005X. doi: https://doi.org/10.1002/wsbm.44. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wsbm.44. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wsbm.44.

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. Z. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, 177(7):1873–1887.e17, June 2019. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2019.05.006. URL https://www.cell.com/cell/abstract/S0092-8674(19)30504-5. Publisher: Elsevier.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, January 2017. ISSN 2041-1723. doi: 10.1038/ncomms14049. URL https://www.nature.com/articles/ncomms14049. Number: 1 Publisher: Nature Publishing Group.