Manifold learning analysis reveals the functional genomics at the cell-type level for neuronal electrophysiology in the mouse brain

Jiawei Huang[1], Daifeng Wang[2,3,#]
[1]Department of Statistics, University of Wisconsin - Madison, Madison, WI, 53706, USA
[2]Department of Biostatistics and Medical Informatics, University of Wisconsin – Madison, Madison, WI, 53706, USA
[3]Waisman Center, University of Wisconsin – Madison, Madison, WI, 53705, USA
[#]corresponding author

Abstract

Recent single-cell multi-modal data reveal different characteristics of single cells, such as transcriptomics, morphology, and electrophysiology. However, our understanding of functional genomics and gene regulation leading to the cellular characteristics remains elusive. To address this, we used emerging manifold learning to align gene expression and electrophysiological data of single neuronal cells in the mouse brain. After manifold alignment, the cell clusters highly correspond to transcriptomic and morphological cell-types, suggesting a strong nonlinear linkage between gene expression and electrophysiology at the cell-type level. Additional functional enrichment and gene regulatory network analyses revealed potential novel molecular mechanistic insights from genes to electrophysiology at cellular resolution.

Introduction

Recent single-cell technologies have generated a great deal of excitement and interest in studying functional genomics at a cellular resolution [1]. For example, recent Patch-seq techniques enable measuring individual cells' multiple characteristics, including transcriptomics, morphology, and electrophysiology in the complex brains, also known as single-cell multi-modal data [2]. Further computational analyses have clustered cells into many cell types for each modality. The same type's cells share similar characteristics: t-type by transcriptomics and e-type by electrophysiology. Those cell types build a foundation for uncovering cellular functions, structures, and behaviors at different scales. However, understanding the molecular mechanisms underlying those linkages is still challenging. In particular, gene regulatory networks (GRNs) connecting the regulatory factors and their target genes, derived from transcriptomic data, can be employed as robust systems to infer genomic functions [3]. Many computational methods have been developed to predict the transcriptomic cell-type GRNs using single-cell genomic data such as scRNA-seq [4]. Primarily, relatively little is known about how genes function and work together in GRNs to drive cross-modal characteristics (e.g., from t-type to e-type).

Also, integrating and analyzing heterogeneous, large-scale single-cell datasets remains challenging. Machine learning has emerged as a powerful tool for single-cell data analysis, such as t-SNE [5], UMAP [6], scPred [7] for identifying transcriptomic cell types. An autoencoder model has recently been used to classify cell types using multi-modal data [8]. However, these studies were limited to building an accurate model as a "black box" and lacked any biological interpretability from the box, especially for cellular phenotypes. To address this challenge, we applied manifold learning, an emerging machine learning field, to align single-cell gene expression and electrophysiological data in the mouse brain. The manifold alignment has better identified many cross-modal cell clusters than existing methods, suggesting a strong nonlinear relationship (manifold structure) linking genes and electrophysiological features at the cell-type level. The enrichment analyses for the cell clusters, including GO terms, KEGG pathways, and gene regulatory networks, further revealed the underlying mechanisms from genes to cellular electrophysiology in the mouse brain.

Results

Manifold learning aligns single-cell multi-modal data and reveals nonlinear relationships between cellular transcriptomics and electrophysiology

We applied a manifold learning analysis to align single-cell multi-modal data for discovering cross-modal cell types (Methods, Fig. 1A). In particular, we aligned 3654 neuronal cells in the mouse visual cortex using their gene expression and electrophysiological data of single cells (two modalities). After alignment, we projected the cells onto a low dimensional latent space and then clustered them into multiple cell clusters. The cells clustered together imply that they share both similar gene expression and electrophysiological features. We have applied multiple machine learning methods to align single cells using two modalities, including linear manifold alignment (LM), nonlinear manifold alignment (NMA), manifold warping (MW), Canonical Correlation Analysis (CCA), and Principal Component Analysis (PCA, no alignment). We found that nonlinear manifold alignment outperforms others (Fig. 1B) based on the Euclidean distances of the same cells on the latent space. This result suggests potentially nonlinear relationships between the transcriptomics and electrophysiology in those neuronal cells, better identified by manifolds. Finally, we visualized the cell alignments of NMA, CCA, and PCA on the 3D latent space in Fig. 1C, showing that nonlinear machine learning has the best alignment (Mean distances of aligned same cells: PCA = 2.117, CCA = 0.510, NMA = 0.132). Besides, we applied our analysis to another multi-modal data of 102 neuronal cells in the mouse visual cortex and also found that the nonlinear manifold alignment outperforms other methods (Fig. S1).

After aligning single cells using multi-modal data, we found that the aligned cells on the latent space by manifold learning recovered the known cell types of a single modality. For instance, those neuronal cells were previously classified into six major transcriptomic types (t-types) based on the expression of marker genes. We also found that the t-types are better formed and recovered by the latent space of NMA than other methods (e.g., CCA and PCA) (Fig. 2A, Fig.

S2). In particular, using the t-types of the cells, we calculated the cells' silhouette values on the latent space after alignment to quantify how well the coordinates of the aligned cells correspond to the t-types (Methods). We found that the silhouette values of NMA are significantly larger than other methods (Fig. 2B), suggesting that NMA better recovers the t-types. Furthermore, NMA revealed a pseudo-timing order across these t-types, implying potential neuronal development aligning with cellular electrophysiology. This developmental trajectory (from Lamp5 to Vip to Serpinf1 to Sncg to Sst to Pvalb) was also supported by previous studies [9]. However, other methods, such as CCA and PCA do not show multiple t-types or trajectories across t-types (Fig. 2A, Fig. S2). Besides t-types, the aligned cells by NMA also revealed morphological types, as shown by aspiny vs. spiny cells in Fig. S3. Thus, these results demonstrate that manifold learning has uncovered known multi-modal cell types from cell alignment.

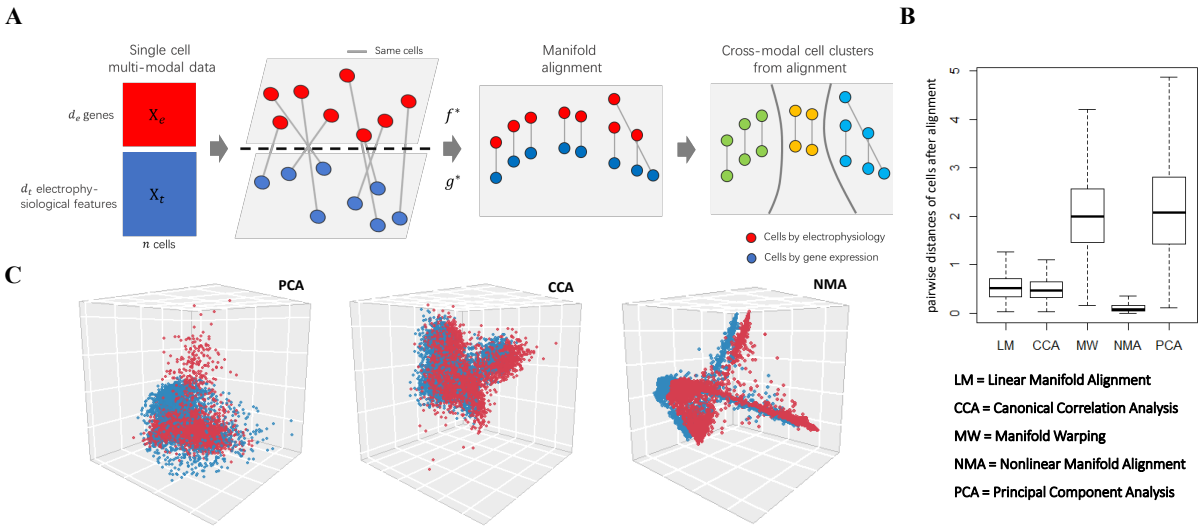

Figure 1 Manifold learning aligns single-cell multi-modal data and reveals nonlinear relationships between cellular transcriptomics and electrophysiology. (**A**) Manifold learning analysis inputs single-cell multi-modal data: X_e, the electrophysiological data (red, d_e electrophysiological features by n cells) and X_t, the gene expression data (blue, d_t genes by n cells). It then aims to find the optimal functions $f^*(.)$ and $g^*(.)$ to project X_e and X_t into the same latent space with dimension d. Thus, it reduces the dimensions of multi-modal data of n single cells to $\tilde{X}_e$ (d reduced electrophysiological features by n cells) and $\tilde{X}_t$ (d reduced gene expression features by n cells). Also, if manifold learning is used, then the latent space aims to preserve the manifold structures among cells from each modality; i.e., manifold alignment. Finally, it clusters the cells on the latent space to identify cross-modal cell clusters. (**B**) Boxplots show the pairwise cell distance (Euclidean Distance) after alignment on the latent space for 3654 neuronal cells in the mouse visual cortex (Methods). The cell coordinates on the latent space are standardized per cell (i.e., each row of $\tilde{X} = [\tilde{X}_e, \tilde{X}_t]$) for comparison across methods. Each box represents one alignment method. The box indicates the lower and upper quantiles of the data, with a horizontal line at the median, the vertical line extended from the boxplot shows 1.5 interquartile range beyond the 75th percentile or 25th percentile. The

machine learning methods for alignment include linear manifold alignment (LM), nonlinear manifold alignment (NMA), manifold warping (MW), Canonical Correlation Analysis (CCA), and Principal Component Analysis (PCA, no alignment). (**C**) The cells on the latent space (3D) after alignment by PCA (no alignment), CCA and NMA. (**D**). The red and blue dots represent the cells from gene expression and electrophysiological data, respectively. The blue dots are drifted -0.05 on the y-axis to show the alignment.

Cross-modal cell clusters by manifold alignment reveal genomic functions and gene regulatory networks for neuronal electrophysiology

Finally, we want to systematically understand underlying functional genomics and molecular mechanisms for cellular electrophysiology using aligned cells. To this end, we clustered aligned cells on the latent space of NMA without using any prior cell-type information. In particular, we used the gaussian mixture model (GMM) to obtain five cell clusters with optimal BIC criterion (Methods, Fig. S4). Those cell clusters are cross-model clusters since they are formed after aligning their gene expression and electrophysiological data. As expected, they are highly in accordance with t-types (Fig. S5). For example, Cluster 4 has ~83.3% Lamp5-type cells (373/448 cells), Cluster 2 has ~77.6% Pvalb-type cells (558/719 cells), Cluster 3 has ~86.6% Sncg-type cells (1339/1546 cells) and Cluster 1 has ~79.1% Vip cells (541/684 cells). Besides, Clusters 1 and 5 include ~55.8% Serpinf1 cells (24/43) and ~60.7% Sncg cells (84/214), respectively. Also, we identified differentially expressed genes (DEGs) with adjusted p-value <0.01 as marker genes of cross-modal cell clusters (Fig. 2C, Supplemental File 1). In total, there are 300, 342, 260, 303, and 22 marker genes in Clusters 1, 2, 3, 4, 5, respectively. These cell-cluster marker genes are also enriched with biological functions and pathways (GO terms) among the genes (Supplemental File 2) (Methods). For example, we found that many neuronal pathways and functions are significantly enriched in DEGs of Cluster 1, such as the ion channel, synaptic and postsynaptic membrane, neurotransmitter, neuroactive ligand receptor, and cell adhesion (adjusted p<0.05, Fig. 2D). Further, we linked top enriched functions and pathways of each cross-modal cell cluster to its representative electrophysiological features (Fig. S6), providing potential novel molecular mechanistic insights for cellular electrophysiology. Also, we predicted the gene regulatory networks for cross-modal clusters that link transcription factors (TFs) to the cluster's genes (Methods, Supplemental File 3), suggesting the regulatory mechanisms for the electrophysiological features in each cluster. For instance, we found that several key TFs on neuronal and intellectual development regulate the genes in Cluster 1, such as Tcf12 and Rora (Fig. 2E). Also, it is interesting to find a subnetwork involving inflammatory TFs and genes such as Irf5 and Spi1 in the network, suggesting potential interactions between neurotransmission and inflammation, which were recently reported [10].

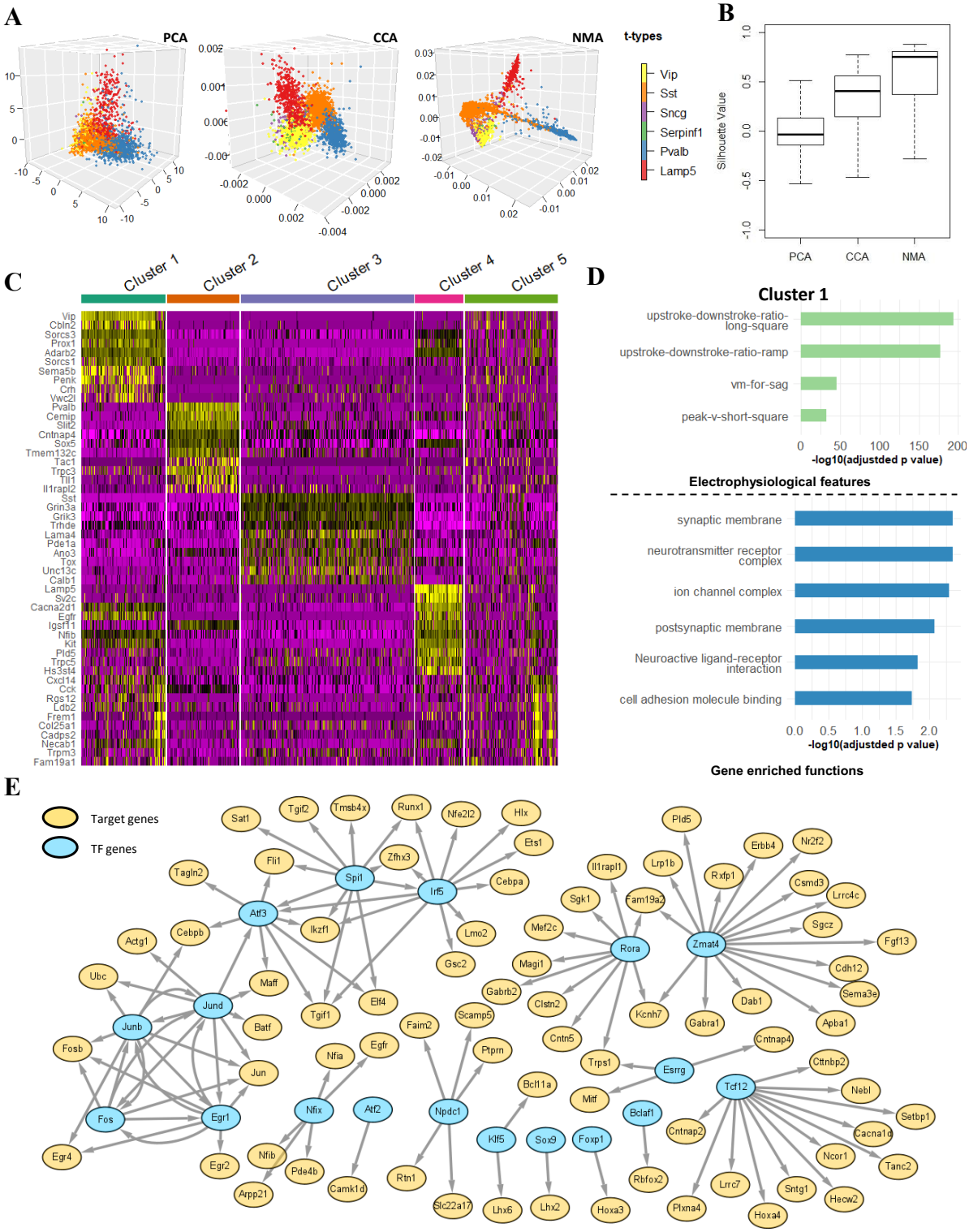

Figure 2 Cross-modal cell clusters by manifold alignment reveal genomic functions and gene regulatory networks for neuronal electrophysiology. (**A**) Scatterplots show 3645 neuronal cells in the mouse visual cortex from electrophysiological data on the latent spaces (3D) after alignment by PCA (no alignment), CCA and NMA. The cells are colored by prior

known transcriptomic types (t-types). Yellow: Vip type; Orange: Sst type; Purple: Sncg type; Green: Serpinf1 type; Blue: Pvalb type; Red: Lamp5 type. The cells from gene expression data on the latent spaces were shown in supplement Fig 3. (**B**) The box plots show the silhouette values of cells for quantifying how well the coordinates of the cells on the latent spaces correspond to the t-types by PCA, CCA and NMA (Methods). (**C**) The gene expression levels across all 3654 cells for Top 10 differential expressed genes (DEGs) of each cross-modal cell cluster. The cell clusters were identified by the gaussian mixture model (Methods). (**D**) The select enriched biological functions and pathways of DEGs (GO and KEGG terms with adjusted p-value <0.05) and representative electrophysiological features (adjusted p-value <0.05) in Cluster 1. (**E**) Gene regulatory networks that link transcription factors (TFs, cyan) to target genes (Orange) in Cluster 1.

Conclusion

In this study, we applied manifold learning to integrate and analyze the gene expression and electrophysiological data of single cells in the mouse brain. We found that the cells are well aligned by the two data types and form multiple cell clusters after manifold alignment. These clusters were enriched with neuronal functions and pathways and uncovered additional cellular characteristics, such as morphology and development. Our results suggest great potential of manifold learning to analyze increasing single-cell multi-omics data and understand single-cell functional genomics in the near future. Our manifold learning analysis is general-purpose and enables studying single-cell multi-modal data in the human brain and other contexts [11].

Methods

Single-cell multimodal datasets

We applied our machine learning analysis for two single-cell multimodal datasets. Primarily, we used a Patch-seq dataset that included the transcriptomic and electrophysiological data of 4435 neuronal cells (GABAergic cortical neurons) in the mouse visual cortex [12]. In particular, the electrophysiological data measured multiple hyperpolarizing and depolarizing current injection stimuli and responses of short (3 ms) current pulses, long (1 s) current steps, and slow (25 pA/s) current ramps. The transcriptomic data measured genome-wide gene expression levels of those neuronal cells. Six transcriptomic cell types (t-types) were identified among the cells: Vip, Sst, Sncg, Serpinf1, Pvalb, and Lamp5. Further, morphological information was provided: 4293 aspiny and 142 spiny cells. Also, we also tested our analysis for another Patch-seq dataset in the mouse visual cortex [13]. This dataset includes 102 neuronal cells with electrophysiological data and gene expression data (Fig. S1).

Data processing and feature selection of multi-modal data

For electrophysiology, we first obtained 47 electrophysiological features (e-features) on stimuli and responses, which were identified by Allen Software Development Kit (Allen SDK) and IPFX Python package [14]. Second, we eliminated the features with many missing values such as short_through_t and short_through_v as well as the cells with unobserved features, and finally selected 41 features in all three types of stimuli and responses for 3654 aspiny cells and 118

spiny cells out of the 4435 neuronal cells. Since the spiny cells usually don't contain the t-type information, we will use the 3654 aspiny cells for manifold learning analysis, and together use the 3654 aspiny cells and 118 spiny cells to refer to morphological cell types (m-type). Also, we standardized the feature values across all cells to remove potential scaling effects across features for each feature. The final electrophysiological data matrix is X_e (3654 cells by 41 e-features). We selected 1302 neuronal marker genes [15] and then took the log transformation of their expression levels. The final gene expression data is X_t (3654 cells by 1302 genes).

Manifold learning for aligning single cells using multi-modal data

We applied manifold learning to align single cells using their multimodal data to discover the linkages of genes and electrophysiological features. In particular, the manifold alignment projects the cells from different modalities onto a lower-dimensional common latent space for preserving local nonlinear similarity of cells in each modality (i.e., manifolds). The distances of the same cells on the latent space can quantify the performance of the alignment. Specifically, given n single cells, let $X_e \in \mathbb{R}^{n \times d_1}$ and $X_t \in \mathbb{R}^{n \times d_2}$ represent their electrophysiological and gene expression data where d_1 is the number of electrophysiological features, and d_2 is the number of genes. Also, $x_e^i \in \mathbb{R}^{d_1}$ and $x_t^i \in \mathbb{R}^{d_2}$ are i^{th} row of X_e and X_t, representing the electrophysiological and gene expression data of i^{th} cell. The manifold alignment aims to find optimal projection functions $f^*(.)$ and $g^*(.)$ to map x_e^i , x_t^i onto a common latent space:

$$f^*, g^* = \underset{f,g}{\operatorname{argmin}}(1-\mu)\sum_{i=1}^{n}\sum_{j=1}^{n}||f(x_e^i) - g(x_t^j)||_2^2 W^{i,j}$$
$$+\mu\sum_{i=1}^{n}\sum_{j=1}^{n}||f(x_e^i) - f(x_e^j)||_2^2 W_{X_e}^{i,j}$$
$$+\mu\sum_{i=1}^{n}\sum_{j=1}^{n}||g(x_t^i) - g(x_t^j)||_2^2 W_{X_t}^{i,j}$$

, where $f^*(.)$ and $g^*(.)$ can be either linear or nonlinear mapping functions, the corresponding matrix $W \in \mathbb{R}^{n \times n}$ models cross-modal relationships of cells (i.e., identity matrix here), and the similarity matrices $W_{X_e}, W_{X_t} \in \mathbb{R}^{n \times n}$ model the relationships of the cells in each modality and can be identified by *k*-nearest neighbor graph (*k*NN). We chose the number of nearest neighbors to be 2, while we also tried other numbers, but the relative performance for different numbers didn't change much (Fig. S7). The parameter μ trades off the contribution between the preserving local similarity for each modality and the correspondence of the cross-modal network. We set $\mu = 0.5$. We used our previous ManiNetCluster method [16] to solve this optimization and found the optimal functions using linear and nonlinear methods, including linear manifold alignment, canonical correlation analysis, linear manifold warping, nonlinear manifold alignment, and nonlinear manifold warping. Finally, after alignment, let $\tilde{x}_e^i = f^*(x_e^i) \in \mathbb{R}^d$ and $\tilde{x}_t^i = f^*(x_t^i) \in \mathbb{R}^d$ represent the coordinates of the i_{th} cell on the common latent space (*d*-dimension) and d be 3 in our analysis for visualization.

Identification of cross-modal cell clusters using Gaussian Mixture Model

After alignment, the cells clustered together on the latent space imply that they share similar transcriptomic and electrophysiological features and thus form cross-modal cell types (i.e., te-types). To identify such cross-modal cell types, we clustered the cells on the latent space into the cell clusters using gaussian mixture models (GMM) with *K* mixture components. Given a cell, we assigned it to the component k_0 with the maximum posterior probability:

$$Pr(k_0|\tilde{x}_{et}^i, \lambda) = \frac{w_i g(\tilde{x}_{et}^i|\mu_{k_0}, \Sigma_{k_0})}{\sum_{k=1}^{K} w_k g(\tilde{x}_{et}|\mu_k, \Sigma_k)}$$

, where $\tilde{x}_{et}^i$ is the i^{th} row of a combined feature set $[\tilde{X}_e, \tilde{X}_t]$, $\lambda = \{w_k, \mu_k|\Sigma_k\}$ $k = 1, \dots, K$ are parameters: mixture weights, mean vectors and covariance matrices. Finally, the cells assigned to the same component form a cross-modal cell type. Also, we used the Expectation-maximization algorithm (EM) algorithm with 100 iterations to determine the optimal number of clusters with K=5 (Fig. S4) by Bayesian information criterion (BIC) criterion [17]. K=5 was chosen at which the $BIC = Kln(n) + 2(\hat{L})$ of the model has an approximately constant and insignificant gradient descent through the equation. Silhouette values are used to compare the clustering result [18], which takes value from -1 to 1 for each cell and indicates a more pronouncedly clustered cell as the value increases.

Differentially expressed genes, enrichment analyses, gene regulatory networks, and representative cellular features of cross-modal cell clusters

We used the Seurat to identify differentially expressed genes of each cell cluster and also multiple tests, including Wilcox and ROC, to further identify the marker genes of cell clusters (adjusted p-value < 0.01) [19]. We applied this method to the electrophysiological features (absolute values) to find each cluster's represented e-features. Also, we used the web app, g:Profiler to find the enriched KEGG pathways, GO terms of cell-cluster marker genes, implying underlying biological functions in the cell clusters [20]. Enrichment p-values were adjusted using the Benjamin-Hochberg (B-H) correction. Furthermore, we predicted the gene regulatory networks for cell clusters, linking transcription factors to target marker genes by SCENIC [21]. Those networks provide potentially novel regulatory mechanistic insights for electrophysiology at the cell-type level.

Supplementary information

Supplemental materials – Supplemental figures
Supplemental file 1 – Differentially expressed genes in cell clusters
Supplemental file 2 – Enrichments of differentially expressed genes in cell clusters
Supplemental file 3 – Gene regulatory networks for cell clusters

Author contributions

D.W. conceived and designed the study. J.H. and D.W. analyzed the data and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

None declared.

Acknowledgments

This work was supported by the grants of National Institutes of Health, R01AG067025, R21CA237955 and U01MH116492 to D.W., U54HD090256 to Waisman Center, and the start-up funding for D.W. from the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison.

Reference

1. Eberwine J, Sul J-Y, Bartfai T, Kim J. The promise of single-cell sequencing. Nat Methods. 2014;11:25–7.

2. Gouwens NW, Sorensen SA, Berg J, Lee C, Jarsky T, Ting J, et al. Classification of electrophysiological and morphological neuron types in the mouse visual cortex. Nat Neurosci. 2019;22:1182–95.

3. Schlitt T, Palin K, Rung J, Dietmann S, Lappe M, Ukkonen E, et al. From gene networks to gene function. Genome Res. 2003;13:2568–76.

4. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020;17:147–54.

5. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nat Commun. 2019;10:5416.

6. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2018;

7. Alquicira-Hernandez J, Sathe A, Ji HP, Nguyen Q, Powell JE. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol. 2019;20:264.

8. Gala R, Budzillo A, Baftizadeh F, Miller J, Gouwens N, Arkhipov A, et al. Consistent cross-modal identification of cortical neurons with coupled autoencoders [Internet]. Neuroscience; 2020 Jul. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.06.30.181065

9. Lim L, Mi D, Llorca A, Marín O. Development and Functional Diversification of Cortical Interneurons. Neuron. 2018;100:294–313.

10. Leite JA, Orellana AMM, Kinoshita PF, Mello NP de, Scavone C, Kawamoto EM. Neuroinflammation and Neurotransmission Mechanisms Involved in Neuropsychiatric Disorders. In: Abreu GEA, editor. Mechanisms of Neuroinflammation [Internet]. InTech; 2017 [cited 2020 Nov 4]. Available from: http://www.intechopen.com/books/mechanisms-of-neuroinflammation/neuroinflammation-and-neurotransmission-mechanisms-involved-in-neuropsychiatric-disorders

11. Berg J, Sorensen SA, Ting JT, Miller JA, Chartrand T, Buchin A, et al. Human cortical expansion involves diversification and specialization of supragranular intratelencephalic-projecting neurons. bioRxiv. 2020;2020.03.31.018820.

12. Gouwens NW, Sorensen SA, Baftizadeh F, Budzillo A, Lee BR, Jarsky T, et al. Integrated Morphoelectric and Transcriptomic Classification of Cortical GABAergic Cells. Cell. 2020;183:935-953.e19.

13. Scala F, Kobak D, Shan S, Bernaerts Y, Laturnus S, Cadwell CR, et al. Layer 4 of mouse neocortex differs in cell types and circuit organization between sensory areas. Nat Commun. 2019;10:4174.

14. Intrinsic Physiology Feature Extractor (IPFX) Python package [Internet]. Available from: https://ipfx.readthedocs.io/

15. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018;362.

16. Nguyen ND, Blaby IK, Wang D. ManiNetCluster: a novel manifold learning approach to reveal the functional links between gene networks. BMC Genomics. England; 2019;20:1003.

17. Huang T, Peng H, Zhang K. MODEL SELECTION FOR GAUSSIAN MIXTURE MODELS. Statistica Sinica. Institute of Statistical Science, Academia Sinica; 2017;27:147–69.

18. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53–65.

19. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177:1888-1902.e21.

20. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler—a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 2016;44:W83–9.

21. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14:1083–6.