Distinguishing neurons from non-neurons using single cell RNA-seq data and clustering methods

Abstract

To increase the granularity with which the brain is characterized, scientists must have a robust method for finding cellular subtypes from single cell transcriptomic data. To this end, we compare the clusters produced by four clustering methods. We then use random forests to identify the genes that best differentiate the cells in a cluster from cells in all other clusters. By comparing and analyzing the biological relevance of the clusters and their associated classifier genes, we conclude that clustering methodology can profoundly impact interpretation of the data. We propose next steps for interpreting our data and testing our findings experimentally.

Background

The incredible diversity of cell types in the brain give rise to complex circuit dynamics that serve as the basis for neural computation. For example, subtypes of GABA-producing hippocampal neurons that differ slightly in cellular makeup produce dynamic network behaviors that might underpin important functions like memory [5]. Characterizing neural subtypes would provide a powerful handle for understanding complex brain behaviors, but traditional biological techniques limit the resolution with which cell types may be studied. The work of Sue McConnell suggests that Satb2-expressing cortical neurons, although indistinguishable as a single population in a stain, may be a functionally heterogeneous population [1]. Such subtle differences, undetectable by histology, might be key in understanding complex neurological disorders like autism or schizophrenia.

Two papers from Ben Barres' lab attempted to use transcriptome profiling to characterize neuronal and non-neuronal subtypes in the murine cortex. In the first paper, two populations of cells from the cortices of week-old mice were fluorescently labeled, purified using flow cytometry, and their mRNA characterized with microarray [3]. All cells that were not fluorescently labeled were considered neurons, which suggests that neural data from this study could have been contaminated by other cell types. The second study improved on this method by specifically purifying neurons using L1 neural cell adhesion molecule, and by expanding the genes studied using RNA-Seq instead of microarray [9]. However, both papers were ultimately limited to studying only those populations of cells that could be identified with a fluorescent antibody, thus eliminating the potential to explore cellular subtypes. Furthermore, the papers had in common only 35% of the 500 genes with the highest neuron-specific expression. This suggests that more work must be done to find a set of genes that define membership to the neuron class.

Recent breakthroughs in single cell RNA sequencing (scRNA-Seq) technologies allow researchers to compare individual cellular transcriptomes. So far, only one paper has attempted scRNA-Seq on brain cell populations [8]. For the first time, researchers were able to take an unbiased approach in the characterization of cellular subtypes. By applying a biclustering algorithm developed by their lab, Linnarsson's group was able to identify 9 classes and 47 subclasses of cells. However, their paper did not include experimental verification of their clustering results. It remains to be seen whether the clusters they identified represent real biological phenomena or are simply a byproduct of their clustering algorithm. The datasets produced by single cell sequencing are vast, and no 'gold standard' method of analysis has yet been established. With this project, we will further an understanding of the impact that the analysis pipeline can have on the biological interpretation of scRNA-Seq.

Problem statement

In this project, we aim to explore the impact that clustering methodology has on the biological interpretation of single cell transcriptome data. First, we will compare the output of multiple clustering algorithms on the Linnarsson single cell dataset. We expect that robust biological phenomena will be detected by multiple kinds of clustering algorithms. Next, we will investigate whether methods other than biclustering are capable of identifying new, biologically relevant subclusters. Finally, we will explore whether the top 100 genes whose expression is most important for distinguishing between clusters, as identified by random forests, provide new potential markers for cellular subtypes. We expect these "classifier gene" lists to include both genes whose expression is enhanced in specific cellular subtypes, and lower-expressed genes with higher-order effects.

Dataset. We will analyze a repository of single cell RNA-Seq data drawn from mouse brain available freely on GEO and used previously in a published study of cell types in cortex [8]. Tissue was taken from 69 juvenile mice (33 male, 34 female) from the either the hippocampus (area CA1) or a part of the cortex (SS). In single cell RNA-Seq, individual cells are isolated using a microfluidic device (Fluidigm C1) and their mRNA are extracted, amplified, and sequenced on an Illumina HiSeq 2000. In order to capture the absolute amount of RNA in each cell rather than relative RNA expression, the authors used unique molecular identifiers to make each mRNA molecule distinct within the cell before the amplification step. The result is a dataset of 3005 samples, each representing the transcriptome of a single cell from either the mouse hippocampus or cortex. The files provided by the authors contain counts already aligned to the genome and associated with functional genomic regions designated with HUGO identifiers. Furthermore, the authors provided labels generated by their method for each cell type: CA1 (hippocampal) neuron, SS (cortical) neuron, inhibitory interneuron, and non-neuronal cells.

Methods

Each entry x_{ij} in the m-by-n feature matrix X represented the absolute abundance, or mRNA transcript count, of gene j in cell i. Because the entries in X are absolute abundances, no extra normalization steps were performed on the data. At m = 3005 data points and n = 19972 genes, the size of the dataset made clustering on the raw data difficult using certain techniques.

For *k*-means and consensus clustering, we reduced the dimensionality of the dataset to the top 500 principal components using principal components analysis. For consensus clustering, we further decreased the size of the dataset to a random selection of 500 cells. We also reduced the number of features in t-SNE and PCA by looking only at those genes reported as cortical neuron specific by Barres et al [3].

t-SNE. As a first step, the dataset was analyzed for batch effects, which would result in clustering due to factors other than cell type. t-distributed stochastic neighbor embedding (t-SNE) is a popular technique for reducing the dimension of the data to an easier to visualize state. t-SNE uses a nonlinear approach to group together more similar data points. We clustered the data using t-SNE and then labeled each cell using the known cell tissue origins provided by Zeisel et al [8]. Specifically, we explored how well interneurons, CA1 neurons, SS neurons, and non-neurons separated from each other. We suspected that in the presence of significant batch noise, the t-SNE output would display differently labeled cells mixed into the same clusters. For our application, *n* could be anywhere from a couple hundred to all 19972 genes of interest. t-SNE was performed using Scikit's t-SNE package [6].

PCA. Principal component analysis (PCA) is another popular technique for visualizing higher-dimensional data in two dimensions. When performing PCA, a covariance matrix of the data is first generated. The high-dimension data points are then projected along the two axes corresponding to the two greatest eigenvectors of the covariance matrix. This allows PCA to capture the directions of greatest variance within the data. PCA is also useful in general for reducing the dimensionality of the data, which significantly decreases the computational load required during clustering. PCA was performed using pre-installed packages in R [7] along with a custom MATLAB implementation.

***k*-means.** *k*-means clustering is an unsupervised learning technique for finding trends in unlabeled data. *k*-means clustering was performed using pre-installed packages in R [7]. One of the greatest challenges in performing *k*-means clustering is in choosing an optimal *k*, and this was accomplished using prior intuition and evaluation of clustering quality over a range of values for *k*. Clustering quality was measured using a ratio of mean distance within clusters to mean distance between clusters. PCA was performed prior to performing *k*-means to reduce the high dimensionality of *X*, the feature matrix.

Consensus Clustering. Consensus clustering, as the name suggested, represents the consensus across multiple runs of a clustering algorithm with different initialization (e.g. K-means). By reconciling clustering information from different runs, the algorithm assesses the stability of discovered clusters. The method has been previously applied to identify subtypes in glioblastoma. Consensus average linkage clustering with Pearson's distance matrix for cluster numbers k = 2 to 6 was performed on 500 cells bootstrapped from the 1691 cortical (SS) cells using an R package. We used a subset of cells from the cortex to ensure a reasonable runtime. The proportional change in the area under the cumulative function (CDF) was used to evaluate cluster stability for finding the optimal number of clusters k. This statistic measures the probability increase by having an additional cluster (from k-1 to k), indicating that the optimal number of clusters corresponds to the k with high value of this proportional change in CDF [Monti *et al. Machine Learning* 2003].

Random Forests. After generating cluster labels using the above techniques, we identified the most important genes for each cluster using random forests. Random forests are an ensemble learning method for classifying data points into known classes. Unlike the LASSO technique, random forests will not throw away important genes that are highly correlated with other important genes in order to minimize an L1 regularization term. For optimal computational performance, we used 10 learners, each limited to $n^{1/2}$ maximum features where n is the dimension of a data point. We noticed that with these hyperparameters, different random forest instances generated different feature importances. To account for this variance, we ran random forest 100 times. The importance of a gene was marked by how often it would appear in the top-100 most important features in individual random forest instances. When working with greater than two clusters, binary labels ("1" if a cell was in a the cluster, "0" if not) were first generated for each cluster. This allowed us to capture which genes were most important for each individual cluster, called "classifier genes" here. Random forests were generated using Scikit [6].

Annotation of Gene Lists. The lists of classifier genes were evaluated for biological importance using DAVID, an annotation tool. In theory, the genes important for classifying a group of cells with a specific, unique function should be enriched for the pathways and proteins related to that function. For example, the production of inhibitory GABA neurotransmitter should be a unique characteristic of inhibitory interneurons. To cluster annotation results in DAVID, we used medium stringency. Only functional annotation clustering results with an enrichment score above 2 were considered. Furthermore, only genes with a modified Fisher exact P-Value less than 0.0001 were considered enriched.

Results

No apparent batch effects. t-SNE using all 19972 genes produced the data points shown in Figure 2. As shown in the figure, coloring of cells based on knowledge of tissue origin resulted in an almost-clean separation of neurons and non-neurons. Interneurons, CA1 neurons, and SS neurons clustered together, however. After trimming the feature set from 19972 to the 480 genes enriched in neurons by Barres et al, the interneurons, CA1 neurons, and SS neurons separated more cleanly (Figure 3). Because t-SNE clusters aligned almost perfectly with the clusters produced by Linnarsson et. al., we concluded that there were no unusual clustering patterns indicative of batch effects in the data. Therefore, we did not apply batch effect correction to this dataset. This is consistent with the approach described by Linnarsson et. al. in their paper. Additionally, the fact that the cells separated cleanly without additional modification of the dataset confirmed that we did not need to perform any extra normalization steps, a process that may remove potentially useful information.

PCA captures main cluster results of Linnarsson biclustering. Although running PCA on the full 3005x19972 size data matrix was too computationally intensive, we successfully ran PCA on the 3005x480 data set, which used just the 480 Barres genes (Figure 3). PCA confirms that even without normalization or accounting for batch effects, the cells cluster nicely into non-neurons, interneurons, CA1 neurons, and SS neurons. Figure 3 also shows how non-neurons form a denser cluster than the three neuron subtypes, suggesting that neurons generally show more variance in expression of the 480 Barres genes, as expected.

Gene selection from Linnarsson biclustering labels. Because both t-SNE and PCA confirmed that the labels generated by the Linnarsson group's biclustering analysis form well-separated clusters, we used these labels to determine what genes distinguish individual clusters of the PCA. The classifier genes separating neurons from non-neurons converged on neuron-specific functions, as we expected, but the classifier genes for neuronal subtypes did not converge on readily differentiable subtype-specific pathways.

To discriminate between neurons and non-neurons, we used the top-100 most important genes chosen by random forests, and then we performed PCA on the newly constructed 3005-by-100 feature matrix (Figure 4).

Although all neurons were cleanly separated from non-neurons, random forests did not attempt to classify cells based on neuron subtypes, leading to interneurons, CA1 neurons, and SS neurons all being clustered into the same group. As expected, we observed significant enrichment of genes related to neuronal functions, including neuron projection growth, vesicles, transmission of electrical signals, and synapses. We also saw many genes with glia-specific functions, such as axon ensheathment.

Next, we used random forests on the three neuron subtypes (CA1 neurons, interneurons, and somatosensory neurons) to determine which 100 genes distinguished a subtype from all other subtypes, resulting in three gene lists. This gene selection process was done only on neurons (1628 total cells). The three gene lists intersected as shown in Figure 5.

We ran PCA on the cells using the intersection of the three sets of 100 genes or 227 unique genes, resulting in the the plots shown in Figure 6.

Although non-neurons were not incorporated in the discovery of the neuronal subtypes, the left plot of Figure 6 shows that the 227 selected genes still separated neurons from non-neurons. Compared to the PCA result in Figure 3, the plots in Figure 6 show better separation of the three neuron subtypes.

This separability between the three neuron subtypes was not reflected in the classifier gene pathway enrichment analysis (Table 1). The classifier genes for all three subtypes primarily functioned in the same pathways - e.g., synapse function, neuron projection. Some PCA clusters were classified by genes whose function was consistent with the cluster's Zeisel label (e.g., classifier genes that identify Linnarsson-labeled CA1 glutamatergic cells were implicated in long term potentiation). As additional confirmation that the PCA clusters were identifying biological clusters, each list of classifier genes contained genes that are known to be selectively enriched in the Zeisel labelled tissue type. For example, interneurons express Pnoc more than other cell types, and Pnoc was identified as a classifier gene for the PCA clusters/Zeisel-cluster of cells called “interneurons”.

k-means identifies different neural subtypes than Zeisel biclustering/PCA

PCA was first performed to reduce the dimension of the feature matrix, and the first 500 principal components (PCs) were selected to run *k*-means analysis. The 3005 cells were projected to two dimensional space using the first two principal components (Figure 7), and the cells were shaded based on k-means clustering assignments. We observed that top principal components tend to have higher values than subsequent ones. Since k-means uses a simple distance metric to minimize within-cluster sum-of-squares, Figure 7 shows that there is a clear separation between the two clusters (k=2) projected using the first two PCs.

For *k* = 2, we observed that cluster 2 was primarily composed of non-neuron cells, whereas cluster 1 was composed of 49% of the interneurons, 54% of the pyramidal CA1, and 59% of the pyramidal SS (Table 2). Thus, non-neurons were easily distinguished from other neuronal cells, but *k*-means struggled with the subclassification within neuronal cells (interneurons, pyramidal CA1, and pyramidal SS). This suggests that *k*-means might be using higher order relationships between genes rather than absolute expression abundances to perform clustering. Using DAVID [4], we analyzed classifier genes that putatively separated neurons and non-neurons. As expected, their functions clustered on neural-specific vesicle production, synapses, protein localization and transport, and neural projections. There were also many classifier genes involved in metabolic processes, which differ between neurons and glia, since neurons rely primarily on oxidative metabolism and glia on glycolysis [2]. Of the 100 classifier genes, 43 were in common with the classifier genes produced in the PCA step above. The 57 classifier genes specific to the 2-means analysis were enriched for myelination and dopamine receptor pathway binding, which both are important functions that separate neurons from non-neurons.

For *k* = 4, we analyzed four gene lists, hoping that the 4 clusters would recapitulate the 4 clusters outlined by Zeisel (i.e., CA1, SS, interneuron, and non-neuron). We used DAVID gene function enrichment to guide our labeling of the identified groups (Table 3). Our second group clearly seemed to implicate glia. Its classifying genes had functions enriched for glial functions (e.g., axon ensheathment - Mbp, Serinc5, Scd2) and proteins (e.g. tetraspanin - Cd81, Cd82). It also included an array of genes involved in oxidative phosphorylation, which is seen more often in neurons. This is likely to be a feature output from random forests as oxidative phosphorylation distinguishes neurons and non-neurons. The gene sets for the other three groups were enriched for neural functions not distinguishable through DAVID analysis alone. Filtering for genes that only appeared in a single list did not give rise to cluster-specific enrichment pathways. Furthermore, none of the classifier gene lists contained the population-enriched genes used to identify the clusters in PCA (e.g., Pnoc to identify interneurons).

In another attempt to classify Groups 1, 3, and 4, we looked at the intersection of each group’s classifier genes with the classifier genes described above produced by using Zeisel labels (Table 4). In theory, if 4-means clustered the same cells as the Zeisel labels, then the random forests would produce similar classifier genes for the corresponding clusters.

Unfortunately, all classifier gene sets maintained the same low, equivalent level of overlap with all of the Zeisel labels. Therefore, *k*-means clustering identified different clusters from PCA and Zeisel's biclustering method.

Consensus Clustering identifies different neural subtypes than *k*-means and Zeisel biclustering/PCA. We ran consensus clustering on 500 cells bootstrapped from the 1691 cells extracted from the somatosensory region of the brain uses three different values of *k*, resulting in the clusters shown in Figure 8.

When *k* = 2, non-neurons were predicted in one class, and inter-neurons

We compared how well the four consensus clusters aligned with the four clusters produced by *k*-means and the four major labels identified by Zeisel et al. Once again, the classifier genes for one group (Group 4) were enriched for non-neural functions, and the classifier genes for all other groups were enriched for general neural functions (Table 6). Many of the classifier genes for Group 4 were specific for pathways involving blood, including response to wounding (Cfp, Stab1, Vnn1), blood clotting von Willebrand factors (Wisp2, Col3a1), and blood-abundant annexin (Anxa1, Anxa2, Anxa3). This leads us to think that the "non-neuronal" cluster includes a small portion of blood cells that contaminated the brain samples during the dissection process and were sequenced, and that true glia were included in the "neuron-specific" clusters.

As with the *k*-means output, classifier genes distinguishing consensus clusters do not overlap significantly with classifier genes distinguishing Zeisel clusters (Table 7). This could be because our random selection of 500 cells from the original dataset of 3005 biased our sample and thus changed the performance of random forests. There is significant overlap, however, between classifier genes distinguishing consensus clusters and those distinguishing *k*-means clusters (Table 8). There is no clear one-to-one correspondence between a *k*-means cluster and a consensus cluster; classifier genes of consensus cluster 4 has no overlap with any *k*-means classifier genes, whereas 29-59% of the classifier genes of consensus cluster 3 overlap with three *k*-means clusters for *k* = 4. The genes in common between *k*-means and consensus clusters are enriched for vesicles (e.g. Hsp90aa1), signaling (e.g. Calm3), and neuron-specific oxidative phosphorylation metabolism (e.g. Atp1a3). This data suggests that although the clusters generated by consensus clustering are likely not the same as the clusters generated by *k*-means, there exists a core set of neuron-specific and metabolic pathway genes with highly variable expression that clustering algorithms tend to use for discriminant analysis.

Conclusions. Our analysis indicates that clustering methodology can have a significant impact on the interpretation of single-cell RNA-Seq data (Table 9). Although all of the methods we implemented were able to distinguish between neurons and non-neurons, only PCA and t-SNE, the two implicit clustering methods, were able to recapitulate the 4 neuronal subtypes observed in the Zeisel paper. None of the methods came close to replicating the 47 subtypes Zeisel's group observed using biclustering analysis.

Rather than characterizing each cluster using genes that were more highly expressed within the cluster, we characterized each clusters using genes that were most important for distinguishing the cluster from all other clusters according to random forests. We initially expected that these two methods would converge on the same set of highly expressed genes. This seems to be the case using the labels produced by the Zeisel biclustering method, wherein the classifier genes produced by random forests include genes that are known to be highly expressed in that neural population (e.g. Pnoc). The clusters produced by our other methods have classifier genes that do not contain population-specific, highly expressed genes defined by others [3]. Although this could indicate flaws in our analysis, we consider the possibility that some clustering methods are sensitive to higher-order relationships between genes of low expression or unique combinations of commonly expressed genes. Although our clustering analyses may not have distinguished between cortical and hippocampal tissue, they might have identified other biologically salient groupings, such as cells with minute differences in their metabolic profile.

Assumptions. In this study, we assumed that the values provided by the author represented true mRNA counts for our dataset. In reality, several entries in the data matrix were likely false negatives because single-cell data tend to be noisy. Some transcripts that were present in the original cell may be missed by the sequencing procedure because of the small amount of starting material. We assumed that with a large enough sample of data points (3005), the noise will average out as more highly expressed genes will less likely become false negatives. We also assumed that the findings of Zeisel et. al. were a good approximation for the ‘true’ dataset, since they provided metadata on which part of the cortex cells were taken from.

Future work

Limitations of Approach. At several points, we reduced the size of our dataset to make it more tractable for analysis (e.g. with PCA). Doing so may have removed potentially informative data and skewed our results. We used a number of principal components that seemed to account for most variance and tried to provide a biological justification for other reductions in the number of cells or genes in our dataset.

Each of our clustering methodologies is potentially limited by the way in which it is calculated, thus the need for comparison. For example, *k*-means is liable to favor the use of highly expressed genes to make classification decisions, because it uses the Euclidean norm to calculate cluster membership. Perhaps evaluating the performance of the clustering algorithms using other norms will result in more biologically relevant clusters.

We also made a critical assumption that our data set was large enough for false negatives in the feature matrix to not have a significant impact on our results. By using a more complex Bayesian model, we may be able to alleviate some of this noise by using information from other cells to detect false negatives. For example, if a gene was expressed in 3004 cells but has an abundance of 0 in the last cell, this value is likely a false negative.

Future Directions. It would be useful to compare the results of these clustering methods with others, including unsupervised learning methods where the number of clusters does not have to be predefined, or in which category membership can be hierarchical or more flexibly defined. In one preliminary analysis, we attempted to use Latent Dirichlet Allocation (LDA) to observe how allowing cells to be members of multiple categories affected clustering. Ideally, the categories (called “topics”) inferred by LDA would represent a unique, probabilistic combination of classifier genes, such that membership in multiple categories would result in a difficult-to-interpret gene expression pattern. LDA did not generate any useful clusters from our results. The four topics had nearly the exact same frequency distribution across all genes (Figure 9). One potential future direction would be to attempt to refine this method or understand why it fails when applied to our dataset.

Based on the similarity of the classifier genes used to identify the 4 consensus clusters and the 4 *k*-means clusters, there might be a core set of genes with variable expression (e.g., Atp1a3, Calm1) that certain clustering algorithms favor when discriminating between cell types. Further investigation is needed to understand if these genes can provide biologically relevant distinction between subtypes, or whether they are an artifact of clustering methodology. Quantitative PCR in single cells could explore whether these genes consistently define categories of cells, thereby adding a new and potentially useful schema to the traditional understanding of cell types.

Contributions

Tiffany performed the *k*-means and consensus clustering. Kristin did LDA clustering and the biological interpretation of the project and the classifier genes. Jesse performed the PCA and t-SNE clustering, as well as extracted classifier gene lists using random forests for every clustering methodology. All three contributed to the proposal, update, and write-up.

References

1. Alcamo, Elizabeth A., Laura Chirivella, Marcel Dautzenberg, Gergana Dobreva, Isabel Fariñas, Rudolf Grosschedl, and Susan K. Mcconnell. "Satb2 Regulates Callosal Projection Neuron Identity in the Developing Cerebral Cortex." Neuron 57.3 (2008): 364-77. Web.
2. Bélanger, Mireille, Igor Allaman, and Pierre J. Magistretti. "Brain Energy Metabolism: Focus on Astrocyte-Neuron Metabolic Cooperation." Cell Metabolism 14.6 (2011): 724-38. Web.
3. Cahoy, J. D., B. Emery, A. Kaushal, L. C. Foo, J. L. Zamanian, K. S. Christopherson, Y. Xing, J. L. Lubischer, P. A. Krieg, S. A. Krupenko, W. J. Thompson, and B. A. Barres. "A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function." Journal of Neuroscience 28.1 (2008): 264-78. Web.
4. Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. "Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources." Nature Protocols Nat Protoc 4.1 (2008): 44-57. Web.

5. Klausberger, T., and P. Somogyi. "Neuronal Diversity and Temporal Dynamics: The Unity of Hippocampal Circuit Operations." Science321.5885 (2008): 53-57. Web.
6. Pedregosa, F., Varoquaux, G., Gramfort, A. & Michel, V. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
7. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
8. Zeisel, A., A. B. Munoz-Manchado, S. Codeluppi, P. Lonnerberg, G. La Manno, A. Jureus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson. "Cell Types in the Mouse Cortex and Hippocampus Revealed by Single-cell RNA-seq." Science 347.6226 (2015): 1138-142. Web.
9. Zhang, Y., K. Chen, S. A. Sloan, M. L. Bennett, A. R. Scholze, S. O'keeffe, H. P. Phatnani, P. Guarnieri, C. Caneda, N. Ruderisch, S. Deng, S. A. Liddelow, C. Zhang, R. Daneman, T. Maniatis, B. A. Barres, and J. Q. Wu. "An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex." Journal of Neuroscience 34.36 (2014): 11929-1947. Web.

Appendix

Figure 1. Pipeline for processing the single-cell RNA-seq dataset.

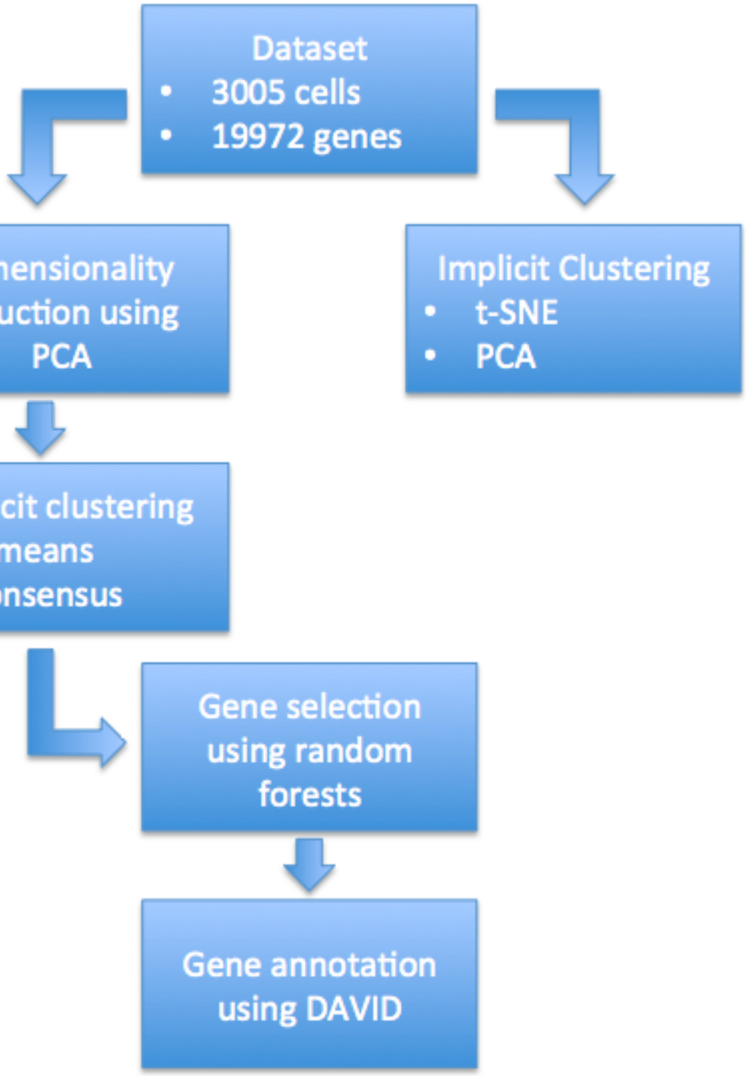

Figure 2. t-SNE results using all 19972 genes.

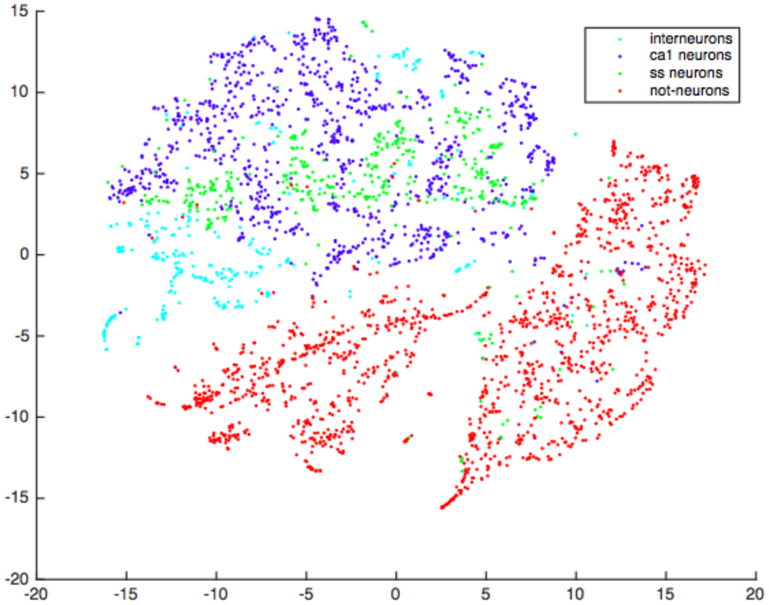

Figure 3. t-SNE (left) and PCA (right) clustering results using 480 genes deemed neuron-specific by Barres et al.

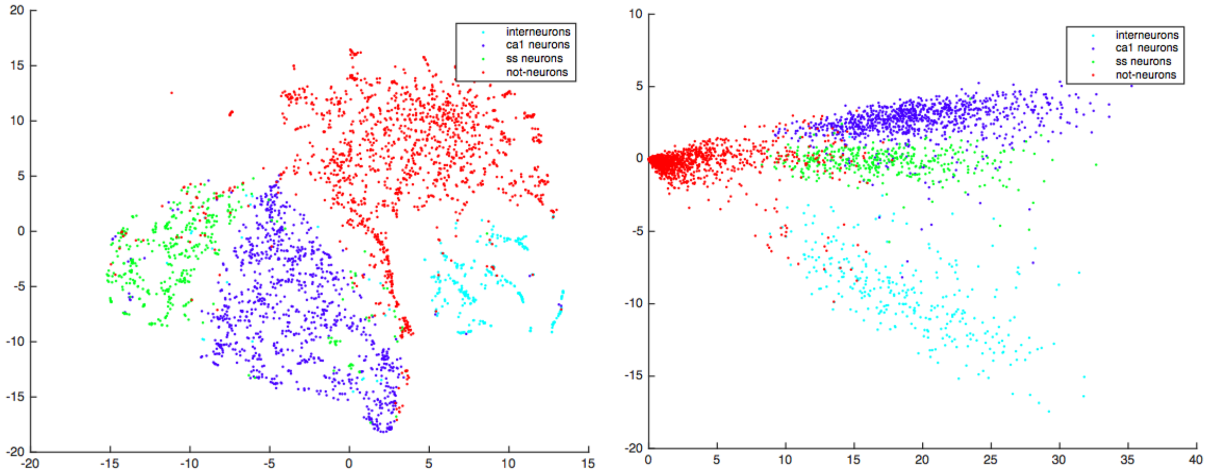

Figure 4. PCA using the 100 most important genes for distinguishing neurons from non-neurons, according to random forests.

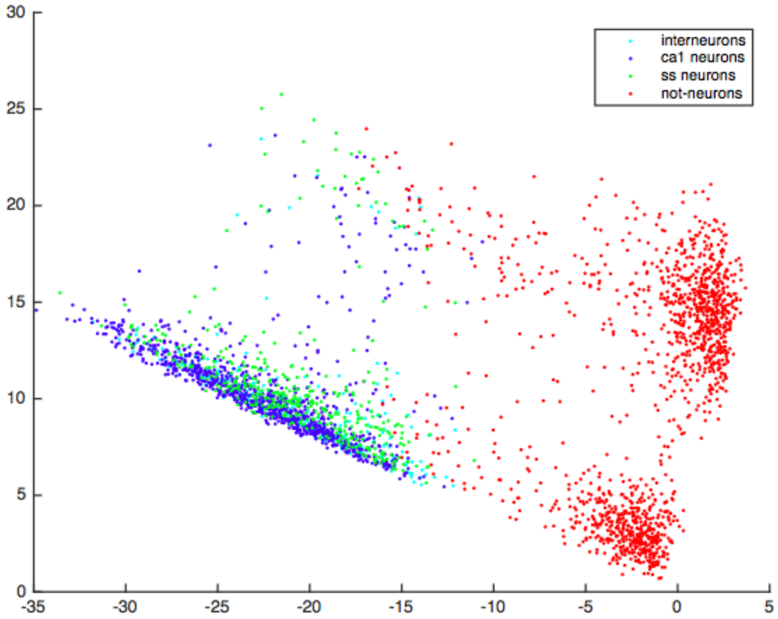

Figure 5. Venn diagram showing how genes selected by random forests for different Zeisel neuron subtypes overlapped.

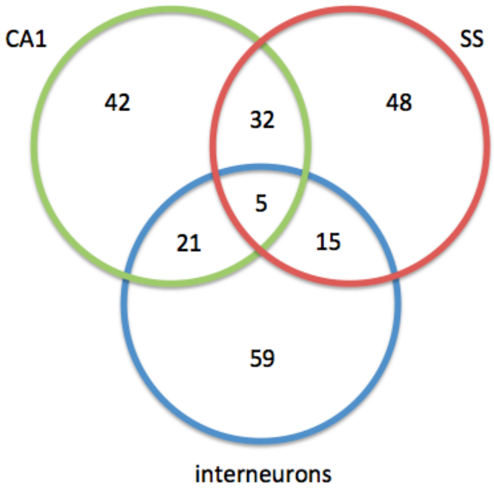

Figure 6. PCA on all cells (left) and just neurons (right) using the 227 most important genes for distinguishing interneurons, CA1 neurons, and SS neurons, according to random forests.

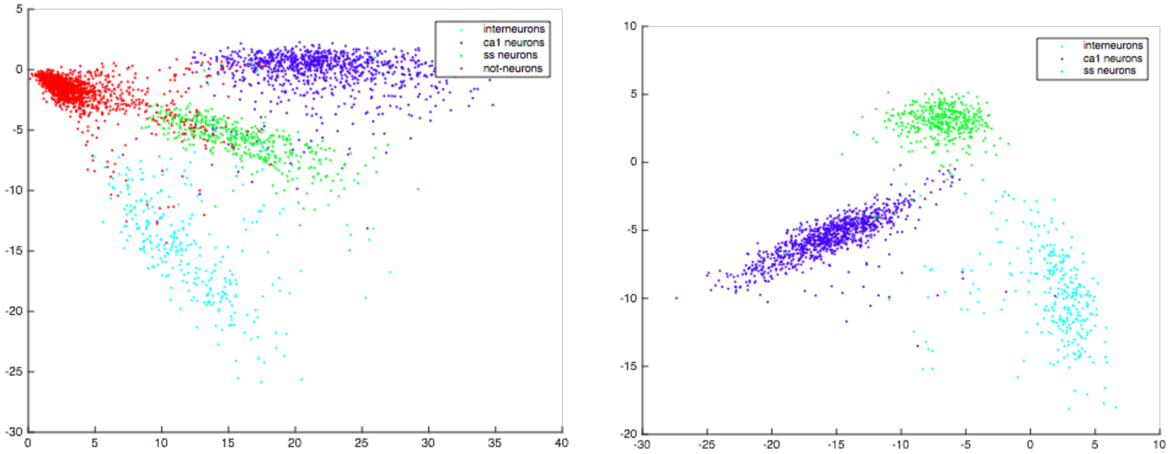

Figure 7. *k*-means clustering results for *k* = 2 (left) and *k* = 4 (right). All 3005 cells were projected to two dimensional space using the first two principal components, and were color-coded based on clustering assignments for k=2 and k=4, respectively.

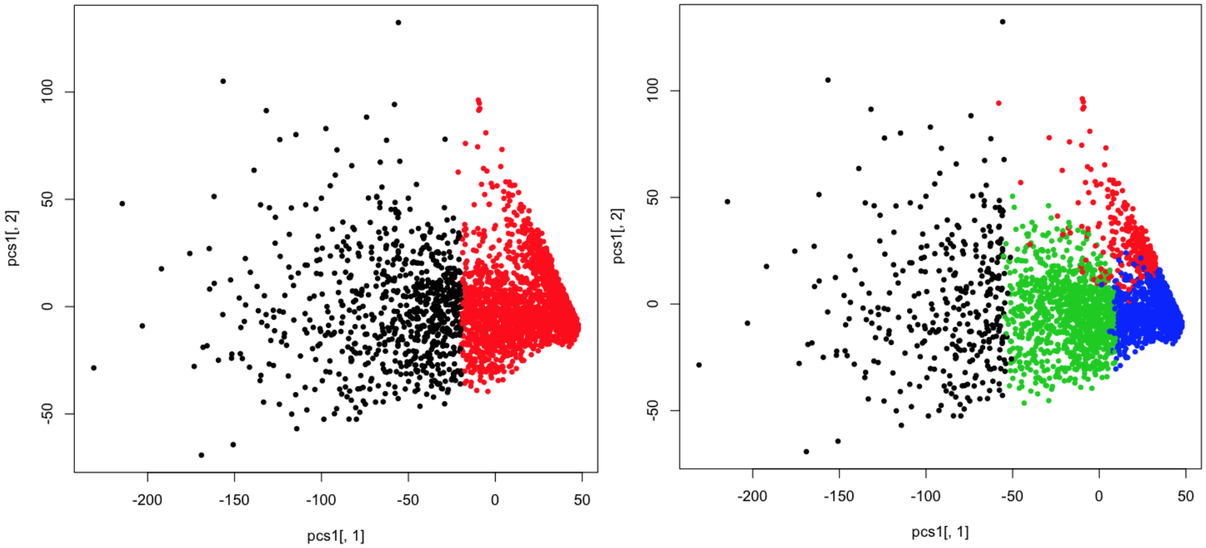

Figure 8. Consensus clustering results for *k* = 2 (first row, left), *k* = 3 (middle), and *k* = 4 (right). Row and column are cells. the relative change in area under CDF curve (second row, left) indicates the optimal number of clusters k=4 .

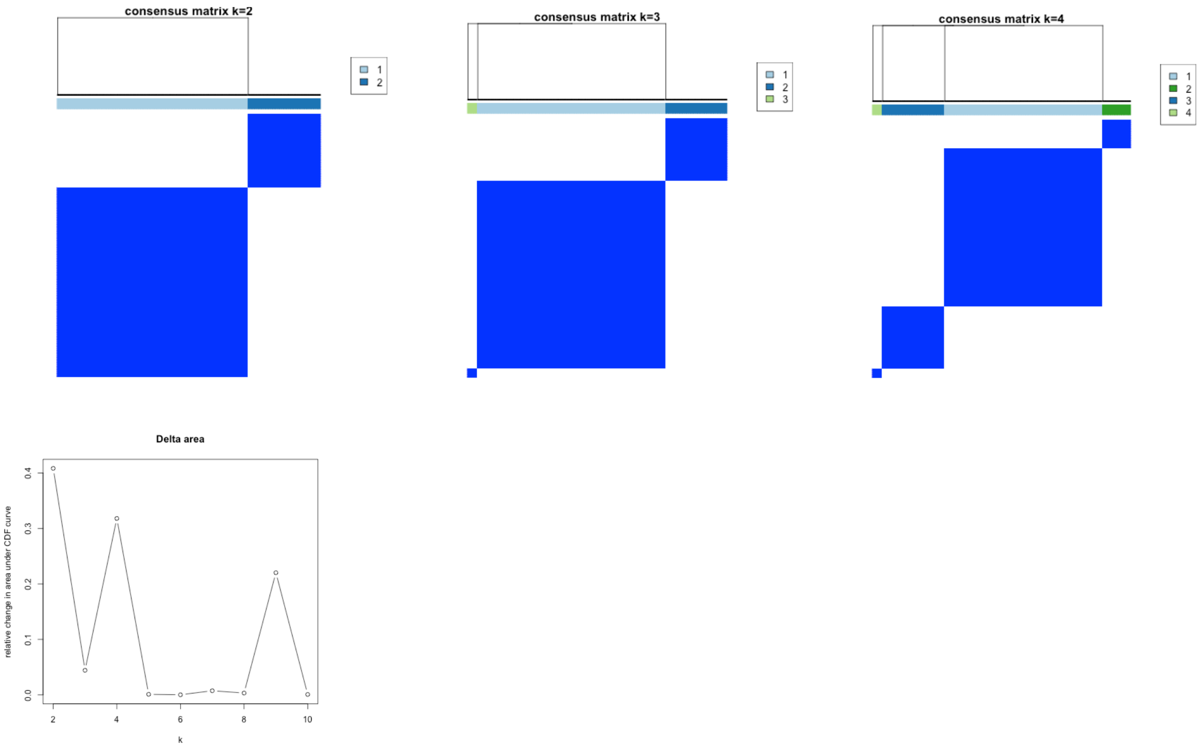

Figure 9. Heatmap of frequency each gene appears in four topics/groups generated with LDA.

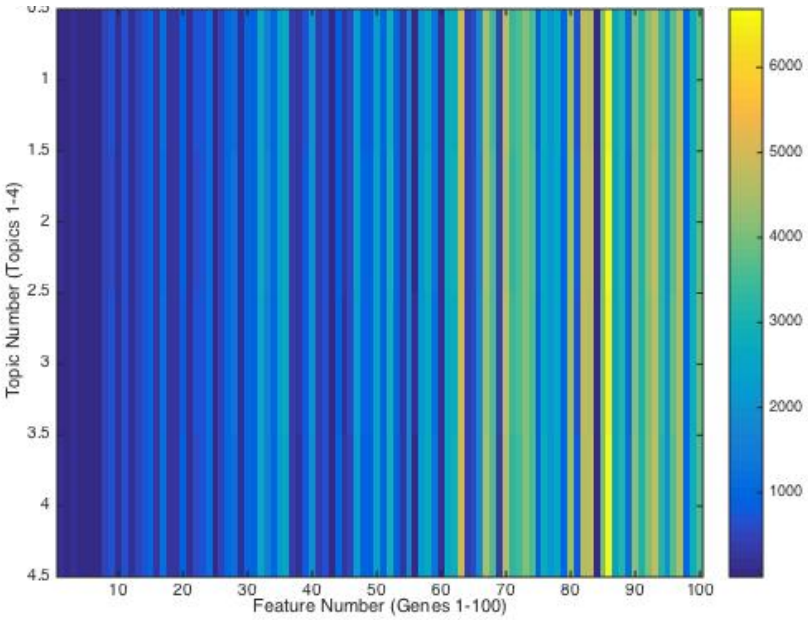

Table 1. Biological implications of genes selected by random forests. Analysis done using DAVID.

Zeisel Label for PCA Cluster	**Pathways relevant to selected genes**	**Subtype-selective genes proposed by Zeisel group that also appeared in RF output**
Interneurons	neuron projection; synapses; neurotransmitter transport; synapses; regulation of neurotransmitter levels; negative regulation of gliogenesis	Pnoc (interneurons); Slc32a1 (interneurons)
CA1 glutamatergic cells	Calcium signaling, transmission of nerve impulse, neural development, cellular homeostasis, GABA receptors, long term potentiation (LTP)	Spink8 (CA1 glutamatergic cells); Fibcd1 (CA1 glutamatergic cells)
SS glutamatergic cells	Neuron projection, neuron development/neurogenesis, synapse, dendrite, behavior/cognition	Gm11549 (SS glutamatergic cells); Tbr1 (cortical projection neurons)

Table 2. Contingency table for *k*-means for *k* = 2 and *k* = 4 using labels from Zeisel et al.

k	**predicted / true**	**interneurons**	**non-neurons**	**CA1**	**SS**
2	1	147	11	435	163
	2	143	1366	504	236
4	1	153	30	623	243
	2	0	301	1	0
	3	61	1041	118	96
	4	76	5	197	60

Table 3. Biological implications of genes selected by random forests on cells labeled using *k*-means clustering (*k* = 4). Analysis done using DAVID.

Cluster	Pathways relevant to selected genes	Suggested Label
1	Neurogenesis functions (Neurotrophin/Cell Cycle/Microtubule Movement/Cell Morphogenesis); Synaptic Transmission; LTP	Neurons
2	myelin sheath/myelination; tetraspanin; oxidative phosphorylation	Non-neurons (glia)
3	oxidative phosphorylation; vesicles; protein transport;	Neurons
4	protein localization; oxidative phosphorylation; vesicle; neuron differentiation; axons; synapse; neurogenesis functions (neuron development/Microtubule movement)	Neurons

Table 4. Intersection of genes selected using Zeisel labels and genes selected using *k*-means labels (*k* = 4). All genes were selected using random forests.

Cluster	Zeisel CA1 cells	Zeisel SS cells	Zeisel Interneurons
1	5	4	4
2	0	2	1
3	1	2	1
4	0	3	0

Table 5. Contingency table for consensus clustering using various values of *k*.

k	predicted / true	interneuron	non-neuron	SS
2	1	5	302	54
	2	36	29	74
3	1	5	302	54
	2	35	13	72
	3	1	16	2
4	1	4	299	2
	2	1	3	52
	3	35	13	72
	4	1	16	2

Table 6. Biological implications of genes selected by random forests on cells labeled using consensus clustering (*k* = 4). Analysis done using DAVID.

Cluster	Pathways relevant to selected genes	Suggested Label
1	Synapse and synaptic transmission, oxidative phosphorylation, regulation of synaptic plasticity, vesicle binding complex proteins, neural projections,	Neuron
2	Oxidative phosphorylation, neuron projection, amyloid-beta precursor protein, synapse and synaptic transmission, locomotion, homeostasis, neurotrophin signaling pathway, ion transportation, axon	Neuron
3	Oxidative phosphorylation, vesicle, synapse and synaptic transmission, neurotrophin signaling pathway, glycolysis, neuron projection,	Neuron
4	proteinaceous extracellular matrix, cell adhesion, response to wounding, von Willebrand factors, annexin, collagen,	Non-neural cell - possibly blood

	tetraspanin	

Table 7. Intersection of genes selected using Zeisel labels and genes selected using consensus clustering labels (*k* = 4). All genes were selected using random forests.

Cluster	**Zeisel CA1 cells**	**Zeisel SS cells**	**Zeisel interneurons**
1	7	6	7
2	7	7	12
3	2	2	4
4	0	0	0

Table 8. Intersection of genes selected using consensus clustering labels (*k* = 4) and genes selecting using *k*-means clustering (*k* = 4). All genes were selected using random forests.

Consensus Cluster	***k*-means cluster 1**	***k*-means cluster 2**	***k*-means cluster 3**	***k*-means cluster 4**
1	53	0	17	39
2	43	12	26	36
3	55	0	29	59
4	0	0	0	0

Table 9. Summary of performances of tested methods.

Method	**Consistent with Zeisel labels**	**N=2 clusters separates neurons from non-neurons**	**N=4 clusters separates non-neurons, CA1, SS, and interneurons**	**Classifier genes similar to genes generated using Zeisel labels**	**Classifier genes similar to genes generated using *k*-means clustering**	**Classifier genes similar to genes generated using consensus clustering**
PCA	Yes*	Yes	Yes*	--	No	No
k-means	No	Yes	No	No	--	Yes
Consensus Clustering	No	Yes	No	No	Yes	--

*but only when filtered for genes specifically enriched in neurons