Clustering cell types in the post-mortem middle temporal gyrus using snRNA-seq data

Thomas Zhang
Computational Biology Department
Carnegie Mellon University
tczhang@andrew.cmu.edu

Sumitra Lele
Computational Biology Department
Carnegie Mellon University
slele@andrew.cmu.edu

Shamieraah Jamal
Computational Biology Department
Carnegie Mellon University
shamierj@andrew.cmu.edu

Jen Yi Wong
Computational Biology Department
Carnegie Mellon University
jenyiw@andrew.cmu.edu

1 Background

The middle temporal gyrus (MTG) is a cortical region in the temporal lobe of the brain which has been implicated in multimodal sensory integration, language processing and even memory [1]. Due to its central role in many cognitive processes, its dysfunction has been found to contribute to various neurological disorders such as schizophrenia [2] and epilepsy [3] through functional magnetic resonance imaging (fMRI) studies. While many mouse models of neurological disorders exist, they do not capture the full etiology of diseases and a stark gap between human and mouse presentations of diseases is still a limiting factor in drug discovery [4]. Therefore, it is desirable to have human tissue samples to understand how the MTG contributes to various cognitive dysfunctions on a molecular level.

Due to the nature of neurological disorders, human brain tissues can only be retrieved through invasive means, such as surgical resections or post-mortem dissections. Indeed, this is how MTG tissues are retrieved for epilepsy, which offers the potential of connecting molecular changes in the brain at various timepoints until post-mortem. However, studying post-mortem tissue often has many challenges, including degradation of cells and poor RNA quality, which is why single nuclei RNA-seq (snRNA-seq) poses a viable option as opposed to single cell RNA-seq (scRNA-seq) [5]. In snRNA-seq, the tissue is first lysed and homogenized in a gentle detergent, separated into single nuclei using fluorescent-activated nuclear sorting (FANS), and then passed through quality control by checking the expression of known housekeeping and marker genes in quantitative polymerase chain reaction (qPCR). Then, the sequences are mapped to a reference genome and further clustering and classifications can be performed [5]. According to Krishnaswami et al. (2016) [5], nuclear transcriptomes largely capture similar variation in cell types to cytoplasmic transcriptomes, although cytoplasmic RNA levels are more indicative of downstream protein expression.

1.1 Objective

In order to facilitate molecular studies of the MTG in post-mortem tissue, this report aims to validate if known cell types in the brain can be detected with unlabelled snRNA-seq data. Firstly, clustering is performed to identify variable groups of nuclei, using either a modularity-based Louvain clustering or a probabilistic Gaussian mixture model method. Secondly, the clusters are annotated by looking for marker genes in each cluster that correspond to known cell types. Finally, a Naïve Bayes classifier is developed to classify new nuclear RNA-seq data using both types of clusterings.

Submitted to Computational Biology Department. 02-620 Machine Learning for Scientists, Spring 2023.

2 Methods

2.1 Dataset

This dataset consists of single nuclear RNA sequencing (snRNA-seq) data from 6 different cortical brain regions. The data is available at: https://portal.brain-map.org/atlases-and-data/rnaseq/human-multiple-cortical-areas-smart-seq. The nuclei were isolated from layers 1, 5 and 6 of the post-mortem human cortex, and consisted of 10% non-neuronal nuclei. Subsequently, the RNA from the isolated nuclei was sequenced and aligned with a reference genome. Expression levels are provided as counts per million (CPM) of both exonic and intronic reads mapped to CRCh38.p2 human reference genome. For our project, we chose to focus on the middle temporal gyrus (MTG) as a case study to develop a pipeline that could be applied to other brain regions. The MTG data consists of 16155 cells (samples) and 50281 genes (features). Prior to classification, the counts will be log-transformed, mean-centered and scaled. This dataset will be partitioned into training and test sets in order to train and test the models, respectively. Hodge et al. had reported the results of clustering the data into cell types (Figure 1, [6]), which can serve as a reference for comparison with our results.

2.2 Principle Components Analysis (PCA)

PCA was used to reduce the dimensionality of the dataset. Due to the high computational time for computing a covariance matrix, this was implemented using `Scikit-learn`'s `PCA` package, with `n_components` of 20, 200 and 2000 attempted. Based on the time taken to compute, 20 components are used for subsequent analyses. The top component of PCA is the vector v that satisfies:

$$\text{argmax}_{|w|=1}\{||Xv||\} = \text{argmax}_{|w|=1}\{\sum_n (x_n \cdot v)^2\}$$

where X is a matrix representing the expression levels for the cells. x_n is the expression levels for one of the cells. Hence, the top 20 components of are the vectors that give the top 20 values when used to evaluate the above equation.

2.3 Gaussian Mixture Model for clustering

The clustering using the Gaussian Mixture Model is an unsupervised method that attempts to fit the data to a mixture of Gaussian distributions. It is similar to the K-Means algorithm, except that the update step uses the Expectation-Maximization algorithm (EM) to fit the parameters of the distributions. To get the initial parameters for our GMM, we chose to use k-NN to assign cells to a cluster, then use those clusters to estimate initial means, covariances and mixture proportions. In the E-step, the probability of each cluster containing a particular cell is represented by:

$$P(x_n|c_i) = (|2\pi\Sigma_i|)^{-\frac{1}{2}} e^{-\frac{1}{2}(x_n-\mu_i)^T\Sigma_i^{-1}(x_n-\mu_i)} \tag{1}$$

Where x_n is a single cell's gene expression levels for all genes. c_i is one of the clusters. Σ is the covariance matrix of the cluster c_i. u_i is the center of the cluster c_i.

Then using equation 1. We calculate the probability of a cell belonging to a cluster.

$$P(c_i|x_n) = \frac{P(x_n|c_i)P(c_i)}{\sum_k P(x_n|c_k)P(c_k)} \tag{2}$$

Then in the M-Step we calculate the means of the new cluster centers as well as the covariance of the new clusters.

The new means are calculated using the formula:

$$\mu = \frac{\sum_n x_n P(c_i|x_n)}{\sum_n P(c_i|x_n)} \tag{3}$$

And the new covariance matrix is calculated using the following.

$$\Sigma = \frac{\sum_n (x_n - u_i)^T (x_n - u_i) P(c_i|x_n)}{\sum_n P(c_i|x_n)} \tag{4}$$

Lastly the new mixture proportions of each cluster are calculated as:

$$P(c_i) = \frac{\sum_n P(c_i|x_n)}{N} \tag{5}$$

The E-step and M-step are repeated iterative, until the log-likelihood stops changing significantly. For our usage, x represents a single sample and its features are the top 20 PCA components.

2.4 Louvain Clustering

The Louvain algorithm is an unsupervised graph-based hierarchical clustering method that starts with each cell as its own cluster in a k-NN graph in feature space then iterates over the edges and aggregates the clusters if the change in cluster membership leads to an overall decrease in modularity, the cost function. Modularity is defined as shown below:

$$Q = \frac{1}{2m} \sum_{i,j} [A_{i,j} - \frac{k_i k_j}{2m}] \delta(c_i, c_j)$$

where $A_{i,j}$ is the weight of the edge between i and j, k_i is the sum of the weights of the nodes attached to node i. c_i is the community of node i. $\delta(c_i, c_j)$ is 1 if $c_i = c_j$ and is 0 otherwise. $m = \frac{1}{2} \sum_{i,j} A_{i,j}$. The update is carried out until there is no more increase in modularity. This method was also used in the referenced paper [6]. More specifically, the change in modularity for each node i for each possible community C is as follows:

$$\Delta Q = \frac{k_{i,in}}{2m} - \frac{k_i \dot{\Sigma}_C}{2m^2}$$

where $k_{i,in}$ is the sum of edges weights of edges from i to nodes in C, k_i is the sum of edges weights of edges from i and Σ_C is the sum of all edge weights of edges from nodes in C and m is the size of the graph.

2.5 Marker gene annotation

Marker genes were found for each cluster using the the following criteria: 1) the marker gene's mean expression in its assigned cluster must be more than two-fold enriched relative to other clusters, 2) the mean expression of the marker gene must be ≥ 1 CPM, and 3) if the cluster has fewer than 10 cells, the cluster is not annotated.

2.6 Multivariate Gaussian Naive Bayes Classifier

The Naive Bayes Classifier is a supervised classifier based on the joint distribution of each feature with each cell type. The benefit of this method is that we do not need to consider all 16155 cells at once. Assuming we have k cell types, we would have a different set of distributions for each cell type. By assuming conditional independence, we further reduce the memory cost to $(k-1) \times 20$ distributions. In implementing this method, we use labels from the previously done clusterings to learn the parameters for each joint distribution. Due to the continuous nature of the gene expression, each joint distribution will be a normal distribution. The probability for each cell type Y is given by:

$$P(y = Y_i|X) \propto \prod_{j=1}^{m} (X_j|Y_i) P(y = Y_i)$$

where X_j is the j-th feature, m is the number of features and Y_i is the i-th cell type. Furthermore, the probabilistic nature of this method allows us to quantify certainty via entropy will tell us how confident the classifier is in its prediction. This will also tell us how well the classification has worked.

2.7 Evaluation

For the clustering, we can evaluate the clusters based on Dunn's index (DI), which involves calculating the ratio of the smallest inter-cluster distance to the largest intra-cluster distance. A larger DI indicates more distinct clustering. We can also use the Silhouette Coefficient (SC) to measure cluster similarity:

$$s = \frac{\bar{x}_{im} - \bar{x}_{im,n}}{max(\bar{x}_{im}, \bar{x}_{im,n})}$$

whereby $\bar{x}_{im}$ is the mean distance between all the points in a cluster m and the centroid of that cluster. $\bar{x}_{im,n}$ is the mean distance between a data point in cluster m and every other cluster centroid n. The SC ranges from -1 to +1, with a higher, more positive score indicating more dense clustering. We can take the mean SC for every data point and plot it for every cluster for comparison.

As for classification, we can construct a confusion matrix to record true positive (TP), false positive (FP), true negative (TN) and false negative (FN). This was done by using Scikit-learn to first binarize all the labels for a given cluster, and then using a One-vs-Rest approach which aggregates all other clusters as the "negative" label. We can then plot the True Positive Rate (TPR) against the False Positive Rate (FPR) to visualize the ROC curve. TPR is defined as $\frac{TP}{TP+FN}$ while FPR is defined as $\frac{FP}{TN+FP}$. A higher area under the ROC curve indicates a better performance.

3 Results and Discussion

3.1 Dimensionality reduction of the dataset was a necessary step.

Given that 50,281 genes were captured in the original dataset, and many genes could have zero values or be correlated with other genes, dimensionality reduction was necessary to capture significantly variable components. Thus, principle component analysis (PCA) was used to identify the number of components needed to capture non-redundant features. The results of using 20, 200 or 2000 components are shown in Figure 1. It was found that the variance captured by each run was 7.04%, 12.2% and 33.2% respectively. Although the variance obtained was not high, running the PCA with more components would have become prohibitively time-consuming. Furthermore, we can see from Figure 1 that the variance contributed by the components beyond the first 4 do not vary very much, with each contributing only a small portion of the overall variance. Thus, it was reasonable to reduce computational time and use the most variable components by using the dataset reduced to 20 principal components.

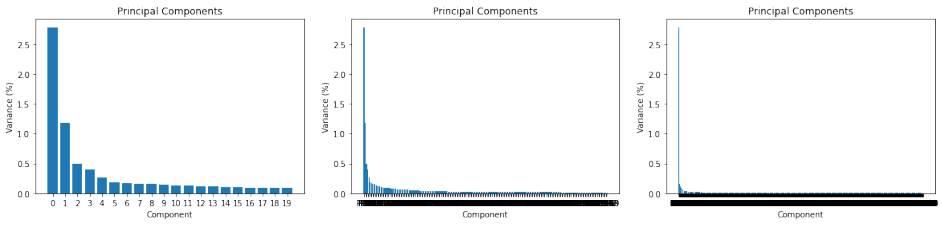

Figure 1: Variance captured by PCA using `n_components` set at 20, 200 or 2000 (left to right).

3.2 Louvain clustering yielded 23 distinct clusters.

A k-NN graph with $k = 5$ was constructed and used for Louvain clustering. The clustering resulted in 53 clusters, of which 30 clusters had less than 10 cells. These clusters were disregarded as they were probably due to noise. Figure 2 shows the results of the Louvain clustering projected on a UMAP. While many of the clusters are clearly defined, there are several clusters that are not cleanly partitioned in the UMAP.

3.3 GMM clustering performs better than Louvain.

We used GMM as a comparison method to Louvain clustering. Since Louvain clustering resulted in 23 clusters, 23 centers were used to cluster with GMM as well. Visually the Louvain clusters seem to be more distinct than the GMM clusters (Figure 2). However the visual trend is not supported by further analysis using Dunn Index and Silhouette Coefficient. We hypothesize that the discrepancy between the appearance of the clusters and the evaluation metrics may be due to the fact that the clustering was performed on the top 20 PCA dimensions. However, when plotting the clusters, we are only taking the top two dimensions.

To analyze the quality of the two clustering methods we calculated the Silhouette Coefficient as well as the Dunn Index (Figure 3). We observed that the clusters made using GMM had slightly higher Dunn index and Silhouette Coefficient than the clusters made using Louvain clusters, indicating that intra-cluster distance was marginally tighter in GMM when compared to Louvain clustering. We believe that this may be due to the fact that GMM naturally creates ellipsoid like clusters while Louvain clustering has no such tendencies.

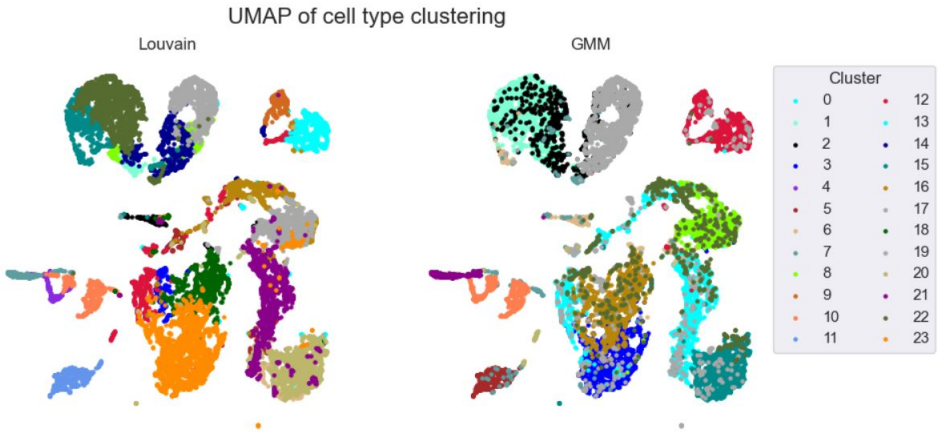

Figure 2: UMAP showing the clustering using Louvain and GMM clustering.

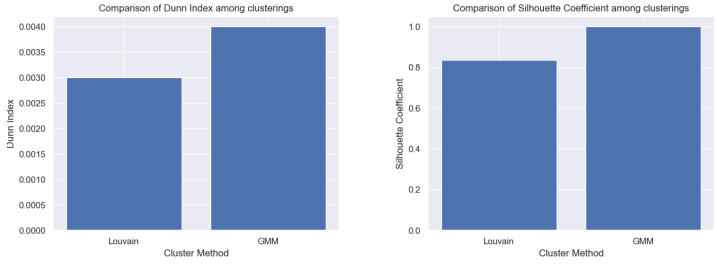

Figure 3: Dunn Index (left) and Silhouette Co-efficient (right) for Louvain and GMM clustering.

3.4 Cell types were annotated using marker genes and literature search.

We used Human Protein Atlas to understand the function and specificity of the marker genes that were identified from our clustering results. We found the genes that had a high mean expression for a given cluster to identify if they could be used as markers for that cluster. Here, we explain the results of the marker genes from the top four clusters.

The first marker gene in cluster 4, ABHD2, has a low tissue specificity but a high cell type specificity and is found in Neurons and Oligodendrocytes. Moreover, ABHD inhibitors have recently been proposed as potential markers for lipid related diseases, and abnormal lipid accumulation is said to be a phenomenon that is a pathological characteristic of epilepsy [7]. The second marker gene, ABCB1, is specifically expressed in enterocytes, which are part of the intestinal lining. However, this gene is associated with the pharmacoresistance in temporal lobe epilepsy[8]. The data that we are looking at does have samples from patients who are being treated for epilepsy hence we hypothesize that this marker gene too, might be relevant.

In cluster 7, the top marker gene A2M is a gene that is strongly associated with Alzheimer's Disease. It is specifically expressed in Microglia and Adipocytes. The second gene ABCC4 is part of the family of ABC transporters which are responsible for transport across the cell membrane and also play a crucial role in regulating the lipid metabolism. RNASE2 is a marker gene enriched in Kupffer cells and is also responsible for lipid metabolism.

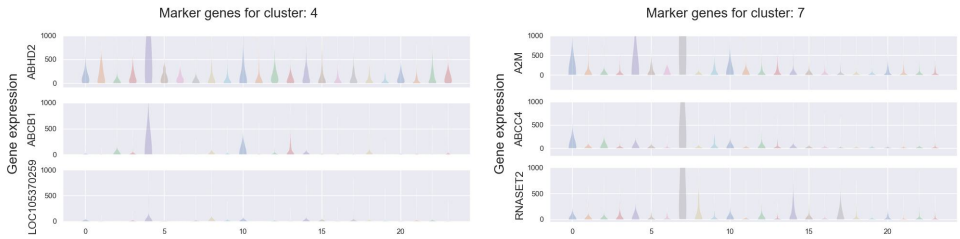

Figure 4: Marker genes found in clusters 4 and 7.

In cluster 10, we found that the marker gene ABLIM1 is specifically enriched in Astrocytes. It has also been found to be differentially expressed in mRNA interactions in epilepsy patients [9]. Lastly, in the top gene in cluster 11, RNF220, is said to be specifically enriched in the brain in Oligodendrocytes.

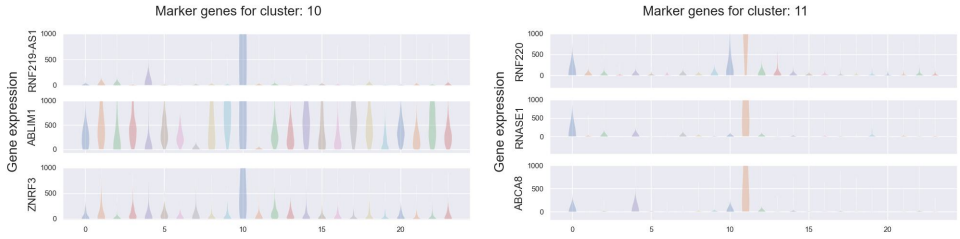

Figure 5: Marker genes found in clusters 10 and 11.

By virtue of finding the correlation between these marker genes and Epilepsy, or their cell-type specificity, we are able to validate our clustering results.

3.5 Naïve Bayes classifier produced good accuracy using both clusterings.

To implement the Naïve Bayes Classifier, we divided the function into three helper functions wherein we initially got the parameters for the data distribution - priors for each class and, mean and standard deviation for each feature - and then used the parameters to fit the test data to the model. We used the parameters obtained from the train data distribution, for each class, for each variable, and then used it in the `fit` function to predict the probabilities for each class for each feature. The train-test split used was 50% train and 50% test. To prevent a bias towards training examples that are from

larger clusters, the sampling method was augmented to allow equal probability of drawing training examples from each class.

We compared the performance of the classifier on data labelled with both clustering methods (Figure 6 and 7). Again, we saw that the performance on the data from the GMM clustering was much better than that of the Louvain clustering. The overall accuracy we got using the labels from GMM clustering was 81.67% and that from the labels from Louvain clustering was 76.77%. This can be seen in the heatmap in Figure 7, which shows much higher concordance of true and predicted label for GMM than Louvain clustering. We can see from the heatmaps that Class 12 in Louvain Clustering tended to misclassified more as Class 3 or 5, whereas Class 8 in GMM clustering tended to be misclassified as Class 14.

This supports the aforementioned possibility that the GMM clustering is superior since the data fits the normal distribution much better. This could also partly be the result of using only 20 PCs as opposed to a much higher value, such that the features used were not sufficient to delineate cell communities more distinctly. Hence, a soft clustering using Gaussian probabilities offers a measure of uncertainty that does not penalize clusters that are not too distinctly separable. As a result, GMM clustering would fit the data that was fed to it better, while it is possible that using more features would allow Louvain clustering to detect communities in a more distinct manner.

On the other hand, the ROC curves in Figure 6 suggest that Louvain clustering yields a better classification, since the area under curve is better for all classes (mean = 0.98) compared to GMM clustering (mean = 0.95). This could simply be due to the fact that more clusters are present in Louvain clustering, with each cluster's membership being smaller than that of GMM, so the probability of misclassification is reduced. This would be more of an inherent feature of the classifier used as opposed to the clustering method, since the classifier would be more familiar with particular members of a given cluster if the cluster size is smaller.

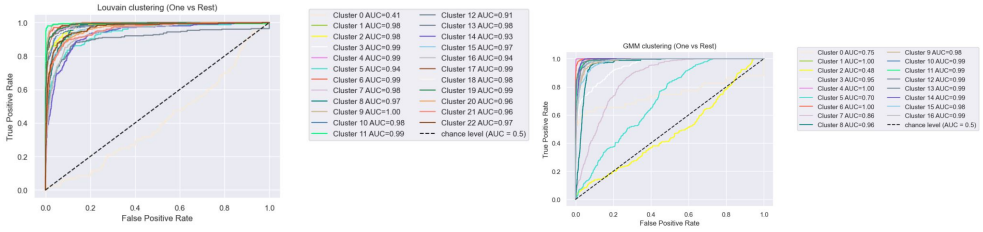

Figure 6: ROC curves for Louvain clustering (left) and GMM clustering (right)

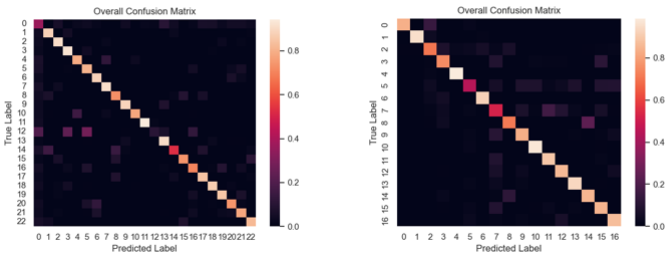

Figure 7: Matrix representing accuracy of Naive Bayes classification on labels from Louvain clustering (left) and GMM clustering (right)

4 Conclusions

Our study further supports that cell types can be differentiated using snRNA seq data. We clustered unlabeled snRNA seq data and were able to identify 23 different clusters. We then labeled the clusters with cell types based on marker genes. With this workflow we managed to identify clusters corresponding to multiple neural cell types including neurons, oligodendrocytes and enterocytes.

One major challenge in our study was determining the number of clusters we should use. More clusters are required for trying to differentiate cell types with more granularity. However, when using more clusters, the clustering methods chosen would begin to create insignificant clusters, (fewer than 10 cells in a cluster).

For the Louvain clustering, some pruning of the edges on the k-NN graph may lead to better clustering. To achieve clusters comparable to the original clustering of the dataset, other post-processing steps would be needed such as merging smaller clusters or clusters with no marker genes.

Another challenge that we faced was underflow. Due to the sparsity of the data, some of the probabilistic calculations would evaluate to a extremely small value. Those small non-zero values would be evaluated as zero, thus we decided to use the implementation of Gaussian mixture from `Scikit-learn` rather than our own because the Gaussian mixture model from `Scikit-learn` uses log-probabilities rather than probabilities and due to time constraints, modifying our own implementation of GMM to use log probabilities was not feasible.

One avenue for future research would be to test the same clustering methods on labeled data. This would help confirm that the cell types determined in this report are accurate. A limitation of this study is that we did not verify our clustering and annotations with existing literature. We used marker genes unique to each cluster to classify the cells and the Dunn index and Silhouette coefficient to evaluate the quality of our clustering. Using labeled data would allow us to annotate the cells with more certainty. In sum, the methods described in this paper were sufficient to obtain cell type classifications from snRNA-seq data, which can be used to identify cells in the post-mortem MTG from nuclear transcriptomes alone for the purpose of facilitating future studies of molecular pathway dysfunction in neuropathologies.

References

[1] Roberto Cabeza and Lars Nyberg. "Imaging cognition II: An empirical review of 275 pet and fmri studies". In: *Journal of Cognitive Neuroscience* 12.1 (2000), pp. 1–47. DOI: 10.1162/08989290051137585.

[2] J.C. Ballenger. "Middle and inferior temporal gyrus gray matter volume abnormalities in first-episode schizophrenia: An MRI study". In: *Yearbook of Psychiatry and Applied Mental Health* 2008 (2008), pp. 260–261. DOI: 10.1016/s0084-3970(08)70836-3.

[3] Bruno J. Weder et al. "Brain areas involved in medial temporal lobe seizures: A principal component analysis of ICTAL SPECT data". In: *Human Brain Mapping* 27.6 (2006), pp. 520–534. DOI: 10.1002/hbm.20196.

[4] Weili Yang et al. "Genetically modified large animal models for investigating neurodegenerative diseases". In: *Cell amp; Bioscience* 11.1 (2021). DOI: 10.1186/s13578-021-00729-8.

[5] Suguna Rani Krishnaswami et al. "Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons". In: *Nature Protocols* 11.3 (2016), pp. 499–524. DOI: 10.1038/nprot.2016.015.

[6] Rebecca D Hodge et al. "Conserved cell types with divergent features in human versus Mouse Cortex". In: *Nature* (Aug. 2019). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6919571/.

[7] Giulia Bononi et al. "/-Hydrolase Domain (ABHD) Inhibitors as New Potential Therapeutic Options against Lipid-Related Diseases". In: *Journal of Medicinal Chemistry* 12.1 (2021), pp. 1–47. DOI: 10.1021/acs.jmedchem.1c00624.

[8] F Zimprich et al. "Association of an ABCB1 gene haplotype with pharmacoresistance in temporal lobe epilepsy". In: *Neurology* 12.1 (2004), pp. 1–47. DOI: 10.1212/01.wnl.0000141021.42763.f6.

[9] X. Li et al. "Identification and Validation of a Dysregulated miRNA-Associated mRNA Network in Temporal Lobe Epilepsy". In: *BioMed research international* 12.1 (2021), pp. 1–47. DOI: 10.1155/2021/4118216.