Identifying sub-populations of cells in single cell transcriptomic data – a Bayesian mixture model approach to zero-inflation of counts

Tom Wilson[1] and Thomas Thorne[*1]

[1]Department of Computer Science, University of Surrey

[*]Corresponding author: tom.thorne@surrey.ac.uk

Abstract

Single cell RNA-seq data exhibit large numbers of zero count values, that we demonstrate can, for a subset of transcripts, be better modelled by a zero inflated negative binomial distribution. We develop a novel Dirichlet process mixture model which employs both a mixture at the cell level to model multiple cell types, and a mixture of single cell RNA-seq counts at the transcript level to model the transcript specific zero-inflation of counts. It is shown that this approach outperforms previous approaches that applied multinomial distributions to model single cell RNA-seq counts, and also performs better or comparably to existing top performing methods. By taking a Bayesian approach we are able to build interpretable models of expression within clusters, and to quantify uncertainty in cluster assignments. Applied to a publicly available data set of single cell RNA-seq counts of multiple cell types from the mouse cortex and hippocampus, we demonstrate how our approach can be used to distinguish sub-populations of cells as clusters in the data, and to identify gene sets that are indicative of membership of a sub-population. The methodology is implemented as an open source Snakemake pipeline avail-

able from `https://github.com/tt104/scmixture`.

1 Introduction

Single cell transcriptomic data gives us a picture of the gene expression in individual cells as opposed to the measurements averaged over many cells produced by bulk experiments. A particular methodology known as single cell RNA-seq (scRNA-Seq) has grown massively in capacity and popularity in recent years[1]. This approach allows for the heterogeneity in populations of cells to be explored by identifying sub populations of cells based on the measured expression of transcripts in individual cells. This is especially relevant in tissues where it is not physically possible to extract cells of a specific known type, for example in the brain[2, 3], where distinct cell types can be inferred from data generated from a mixture of cells of different types. The challenge in this scenario is to allow cell types within a sample to be identified as an unsupervised learning problem, where we wish to cluster together cells of similar types, based on their expression[4]. Although this can be approached by mapping cells to specific reference expression patterns, there will always be scenarios where typical expression patterns are not available[5], or where cells with atypical expression are of interest, for example in disease studies[3, 6].

Work in the literature has applied a range of different approaches to this problem, some making use of statistical methods, whilst others apply machine learning techniques. Statistical approaches include the use of a multinomial model of counts derived from single cell experiments in a Dirichlet process mixture model[7], and hierarchical Bayesian nonparametric models which allow for clustering of data at both the cell and subject level[8]. These Bayesian nonparametric approaches have the advantage that the number of components in the model is learnt from the data, and we adopt a similar Dirichlet process mixture modelling approach in this work. In Duan et al.[7] the single cell RNA-seq data are modelled as counts following a multinomial distribution, as this cir-

cumvents issues around zero-inflation, but we have found that the multinomial distribution to provide a poor fit to scRNA-seq count data, and that the (zero-inflated) negative binomial distribution produces a significantly better fit (see supplementary materials).

Some approaches, for example Monocle[9], SC3[10], and Seurat[11] apply clustering techniques such as K-means, DBSCAN or Louvain clustering, after applying transformations to the data to produce a representation of each cell more amenable to these approaches[12]. While the clustering itself is then often fast and straightforward to apply, such methods do not attempt to model the statistical distribution of the data, and unlike Bayesian nonparametric methods, often require a trial and error approach to parameter tuning, for example applying the model to each possible number of clusters in the data and applying heuristics to determine the number of clusters in the data.

Several machine learning based approaches applying neural networks using autoencoders or similar embedding approaches have been proposed [13, 14, 15], representing expression data in a lower dimensional vector space, on which clustering using methods such as K-means can be performed. However neural network based approaches such as autoencoders suffer from the drawback that results are not easily reproducible, as multiple autoencoder neural networks trained on the same data will give different embeddings, and there is no clear indication of which are superior. There is also no quantification of uncertainty in cluster assignments using these approaches.

A particular benefit of a statistical approach is in the interpretability of the model, whereby the learned clusters can easily be associated with patterns of gene expression through the model. Although attempts have been made to build interpretable machine learning models, for example applying autoencoders with linear decoders[15], these methods still lack the inclusion of posterior uncertainty provided by Bayesian models, both in terms of cluster assignment and gene expression within clusters, and still suffer from the issues with reproducibility described above.

Recently latent factor models have also been applied to scRNA-seq data[16], and allow us to identify important patterns in gene expression across cells. Although these methods can produce useful biological insights, and they are capable of identifying factors that correspond to individual cell types, the task they perform is not directly related to that of clustering. Rather than identifying patterns of expression that define distinct groups of cells, latent factor models learn a set of basis expression signatures that can be combined in different weights to represent the expression across all cells in the data, and the number of factors to learn must be specified in advance. As a result they are less useful for the specific task of identifying sub-populations of cells within a data set.

Single cell transcriptomic data has been said to suffer from a phenomenon know as dropout. It is hypothesised that this occurs when RNA fragments are missed for reverse transcription into cDNA prior to amplification, and results in large numbers of zero counts in the data, which in turn leads to distributions of counts with large spikes at zeros. There has been some discussion in the literature around the issue of zero inflation in single cell transcriptomic data, and it is argued in Svensson[17] that the proportions of zero counts are no larger than would be expected from a negative binomial distribution. However in Clivio et al.[18] a Bayesian model selection approach demonstrates that this is not the case across all data sets, and that the phenomenon may only be observed in a subset of genes. Furthermore by applying a straightforward model selection approach to counts within known cell types, we find that a zero-inflated negative binomial distribution is a superior model for some genes, and that the appropriate model varies from gene to gene. These results are illustrated in figure 1. We have therefore developed a Bayesian nonparametric clustering approach that employs a mixture of a spike at zero and a negative binomial distribution with a transcript dependent mixing proportion. This allows genes that do exhibit inflated zero counts to be appropriately modelled, and is outlined in the methods selection below.

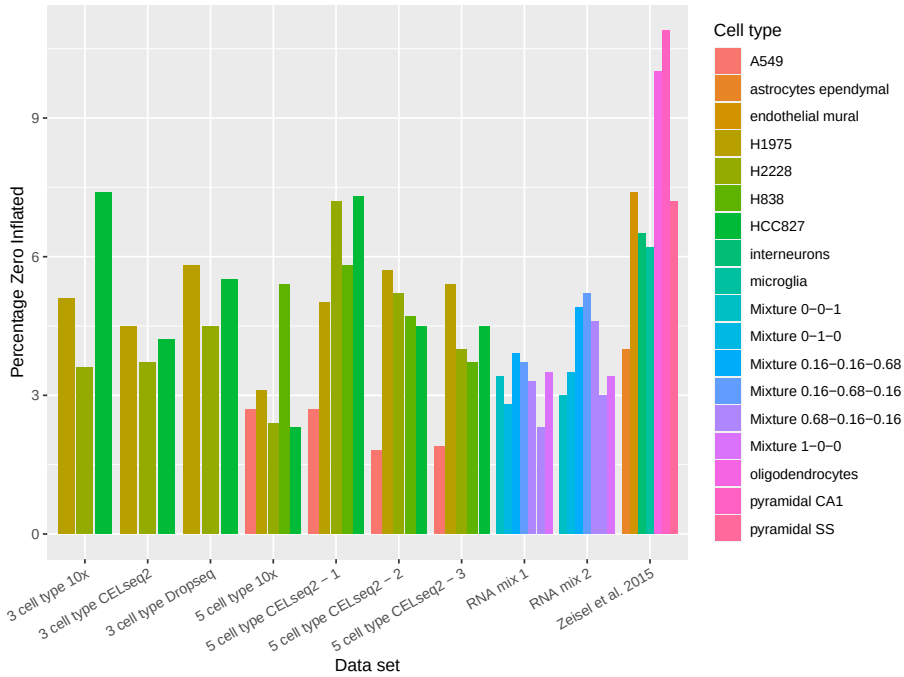


Figure 1: Percentages of genes in each cell type, across data sets from Tian et al.[19] (data sets 1-9) and Zeisel et al. 2015[20], where the zero-inflated negative binomial model is preferred to the negative binomial model.

2 Methods

2.1 Model selection of within-cluster distributions

To investigate the appropriate model for single cell RNA-seq counts within a given cell type, we applied the Bayesian Information Criterion (BIC) given in equation 1 to maximum likelihood parameter estimates $\hat{\theta}$ for individual genes from a subset of 1000 selected as having the largest coefficient of variation, comparing a negative binomial and zero-inflated negative binomial distribution for the likelihood function f,

$$\mathrm{BIC} = k \log n - 2 \log f(\hat{\theta}). \tag{1}$$

Here k represents the number of parameters of the model, and penalises unnecessarily complex models, while n is the number of data points observed. We also considered Poisson, multinomial, negative binomial and zero-inflated negative binomial models across the full set of 1000 genes rather than at the individual gene level, again calculating the BIC as in equation 1.

2.2 scRNA-Seq counts model

For each cell i and each transcript j, counts of reads y_{ij} are modeled as a mixture of an atom at zero, representing dropout, and a negative binomial distribution with mean μ_j and dispersion ω_j

$$y_{ij} \sim w_j \mathrm{NB}(s_i \mu_j, \omega_j) + (1 - w_j)\delta \tag{2}$$

with transcript specific dropout rates w_j. There is also a fixed cell specific scaling factor s_i that compensates for the variation in sequencing depth between cells. Scaling factors can be estimated using existing approaches in the literature, for example scran [21].

We then use Gamma priors on the μ and ω, with the priors determined using an empirical Bayes approach. For each gene j in the data set, we apply

maximum likelihood inference to the model in equation 2, to estimate w_j, μ_j and ω_j, using numerical optimisation. This gives a sample of estimates $\hat{\mu}_j$ and $\hat{\omega}_j$, which are taken to be a sample from the priors on μ and ω respectively, and a Gamma distribution is fit to these using maximum likelihood to find appropriate hyperparameters for our model.

2.3 Bayesian nonparametric clustering model

We introduce a clustering of the data to represent sub-populations of cells by placing a Dirichlet Process distributed prior $\mathrm{DP}(\alpha, H)$ with concentration parameter α on the mean and dispersion parameters of all genes. Here H corresponds to the gamma priors on μ and ω. This can be represented using the stick breaking construction of the Dirichlet process, and sampling performed by introducing a latent variable describing the cluster membership of each cell, z_i. Then we can write:

$$
\begin{aligned}
\theta_k &\sim H \\
u_l &\sim \mathrm{Beta}(1, \alpha) \\
\beta_k &= u_k \prod_{l=1}^{k-1} (1 - u_l) \\
P(z_i = k|\beta) &= \beta_k \\
y_i &\sim f(y_i|\theta_{z_i}).
\end{aligned} \tag{3}
$$

where H is the prior on μ and ω, $\theta_{z_i} = (\mu_{z_i\cdot}, \omega_{z_i\cdot})$, the cluster specific mean and dispersion parameters for each gene, and k and l range over $\mathbb{Z}^+$.

The full model, including priors and combining the mixture at the gene level and mixture of mean and dispersion parameters at the cell level, is then given by

$$
\begin{aligned}
y_{ij} &\sim w_j \mathrm{NB}(s_j \mu_{z_i j}, \omega_{z_i j}) + (1 - w_j)\delta \\
w_j &\sim \mathrm{Beta}(1, 1) \\
\mu_{kj} &\sim \mathrm{Gamma}(a, b) \\
\omega_{kj} &\sim \mathrm{Gamma}(c, d) \\
u_l &\sim \mathrm{Beta}(1, \alpha) \\
\beta_k &= u_k \prod_{l=1}^{k-1} (1 - u_l) \\
P(z_i = k|\beta) &= \beta_k.
\end{aligned}
$$

2.4 Markov Chain Monte Carlo sampling scheme

To sample the cluster allocation variables z_i and cluster specific parameters $\mu_{k\cdot}$ and $\omega_{k\cdot}$, we use the stick breaking sampler described in Liverani et al.[22]. This samples variables of the stick breaking construction u_l through the use of auxiliary variables, and retrospective sampling to sample parameters for unoccupied clusters, and then samples each z_i conditionally on the stick breaking lengths.

To allow us to sample from the posterior on the mean and dispersion parameters of the negative binomial distribution μ and ω we introduce auxiliary variables to model the negative binomial distribution as a Poisson–Gamma mixture. This gives

$$
\begin{aligned}
y_{ij} &\sim w_j \mathrm{Poisson}(s_i \mu_j v_{ij}) + (1 - w_j)\delta_0 \\
v_{ij} &\sim \mathrm{Gamma}(\omega_j, \omega_j).
\end{aligned} \tag{4}
$$

The mixture weights w_j in equation 4 are then replaced with indicator variables t_{ij} so that

$$p(y_{ij}|\mu_j, v_{ij}, t_{ij}) = \mathbb{1}(t_{ij} = 1)\left[e^{-s_j\mu_j v_{ij}}\frac{(s_j\mu_j v_{ij})^{y_{ij}}}{y_{ij}!}\right] + \mathbb{1}(t_{ij} = 0)[\delta(y_{ij})] \quad (5)$$

For each transcript j within a cluster k, there is a cluster specific mean μ_{kj}, with prior Gamma(a, b) that is sampled conditionally as

$$\begin{aligned} p(\mu_{kj}|v_{\cdot j}, y_{\cdot j}, a, b, z) &\propto p(y_{\cdot j}|\mu_{kj}, s, v_{\cdot j})p(\mu_{kj}|a, b) \\ &\propto \prod_{i|z_i=k}\left[\frac{(s_i\mu_{kj}v_{ij})^{y_{ij}}e^{-s_i\mu_{kj}v_{ij}}}{y_{ij}!}\right]^{\mathbb{1}(t_{ij}=1)} \frac{b^a}{\Gamma(a)}\mu_{kj}^{a-1}e^{-\mu_{kj}b} \\ \mu_{kj}|v_{\cdot j}, y_{\cdot j}, a, b, z &\sim \text{Gamma}\left(a + \sum_{i|z_i=k} y_{ij}t_{ij}, b + \sum_{i|z_i=k} s_i v_{ij} t_{ij}\right) \end{aligned} \quad (6)$$

The indicator $t_{ij}$is sampled from a multinomial where

$$p(t_{ij} = 1) = \frac{p(y_{ij}|\mu_{z_i j}, s_i, v_{ij})}{p(y_{ij}|\mu_{z_i j}, s_i, v_{ij}) + \mathbb{1}(y_{ij} = 0)}. \quad (7)$$

Then given $t_{ij} = 1$ we can sample v_{ij} as needed, for i s.t. $z_i = k$ using

$$\begin{aligned} p(v_{ij}|\mu_{z_i j}, y_{ij}, s_i, \omega_{z_i j}, z, t_{ij} = 1) &\propto p(y_{ij}|\mu_{kj}, s_i, v_{ij})p(v_{ij}|\omega_{kj}) \\ &\propto \left[\frac{(s_i\mu_{z_i j}v_{ij})^{y_{ij}}e^{-s_i\mu_{z_i j}v_{ij}}}{y_{ij}!}\right] v_{ij}^{\omega_{z_i j}-1}e^{-v_{ij}\omega_{z_i j}} \\ v_{ij}|\mu_{kj}, y_{ij}, s_i, \omega_{kj}, t_{ij} &\sim \text{Gamma}(y_{ij} + \omega_{kj}, s_i\mu_{kj} + \omega_{kj}). \end{aligned} \quad (8)$$

To sample the dispersion parameter ω_{kj} with prior Gamma(c, d) we apply a slice sampler [23] on $p(\omega_{kj}|v_{\cdot j}, c, d, t)$ for each j independently:

$$\begin{aligned} p(\omega_{kj}|v_{\cdot j}, c, d, z) &\propto \prod_{i|z_i=k} p(v_{i,j}|\omega_{kj})^{\mathbb{1}(t_{ij}=1)}p(\omega_{kj}|c, d) \\ &\propto \prod_{i|z_i=k}\left(\frac{\omega_{kj}^{\omega_{kj}}}{\Gamma(\omega_{kj})}v_{ij}^{\omega_{kj}-1}e^{-v_{ij}\omega_{kj}}\right)^{\mathbb{1}(t_{ij}=1)} \frac{d^c}{\Gamma(c)}\omega_{kj}^{c-1}e^{-\omega_{kj}d} \end{aligned} \quad (9)$$

The MCMC sampler iterates over a loop that first updates the within cluster specific parameters $\mu_{k\cdot}$, $\omega_{k\cdot}$, sampling the appropriate v_{ij} as needed to do so, then updates the cluster allocation variables z_i as described in detail in Liverani et al.[22], and finally samples the global indicators t_{ij}.

3 Results

3.1 Model selection for scRNA-seq count data

Using the model selection approach described above, we considered the best models for count data from scRNA-seq benchmarking experimental data[19] as well as an scRNA-Seq data set from the mouse cortex and hippocampus taken from the literature[20].

Results are presented in figure 1, showing the percentage of genes in each cell type within an experiment that were found to be better explained by the zero-inflated model. It is apparent that there are zero-inflated genes across each data set and cell type, and although the percentages are low, they are consistent across experimental platforms. This suggests that using a model that is able to include zero-inflation at different degrees across each gene is a more realistic model of real world scRNA-seq data than using a negative binomial model alone.

Considering models applied across all genes, the results in table 1 show that the negative binomial model is generally superior when applied in this way, except for in the case of the RNA mixture from bulk cell populations, in which the multinomial is preferred. This is perhaps to be expected as the construction of individual cell samples in this experiment does not reflect the process by which individual cells are sequenced in genuine scRNA-seq data sets.

3.2 Identification of sub-populations in scRNA-Seq data

We applied our clustering method to data from benchmarking experiments[19] as well as an scRNA-Seq data set from the literature[20]. For comparison, we implemented a model where the counts within a cluster follow a multinomial

distribution, in the same Dirichlet process mixture model sampling framework as used in our zero-inflated negative binomial model. Scaling factors for each cell were inferred using the scran[21] R package. As in many scRNA-Seq data analysis methods, to reduce the computational burden we first filtered the set of transcripts to the 1000 with the highest coefficient of variation across all cells, when applying the multinomial and negative binomial models. The cluster allocations in the MCMC samplers were initialised using a K-means clustering of the log transformed data with 40 components. We ran four Markov Chain Monte Carlo (MCMC) chains with burn-in for 1000 iterations of the outer loop of the sampler, and selected the maximum a posteriori (MAP) estimates of the allocation variables z_i across all samples collected. We compare our method to an implementation using a multinomial count distribution, similar to that used in Duan et al.[7], and two of the top performing scRNA-Seq clustering algorithms studied in Tian et al.[19], SC3 [10] and Seurat [11]. Both SC3 and Seurat were run on the log transformed data using default settings.

3.2.1 Single cell benchmarking data

To test the performance of our method on data where there is a known ground truth set of cell populations, we applied our methodology to the single cell benchmarking data sets of Tian et al.[19]. These provide data from various mixtures of human lung adenocarcinoma cell lines, designed for the benchmarking of single cell RNA-Seq analysis pipelines, and were sequenced using a number of different platforms.

The first two benchmarking data sets, the three cell and five cell mixtures, are comprised of mixtures of cells from HCC827, H1975, and H2228 cell lines, and mixtures of cells from HCC827, H1975, H2228, A549, and H838 cell lines respectively. These were sequenced using plate (CEL-seq2 [24]), and droplet based (10x[25] and Drop-seq[26]) platforms.

The different methods considered were evaluated using the Adjusted Rand Index (ARI)[27] using the known cell type labels included in the data, over

multiple repated runs of the methods on the data. For SC3 and Seurat, the results did not vary between runs, and so show no variance across runs. The multinomial and zero-inflated negative binomial models use MCMC methods to sample from a model posterior, and so are inherently subject to some variation in the results between runs.

Results from the three cell mixture are shown in figure 2(a). It can be seen that our negative binomial model outperforms both the multinomial model and SC3 and Seurat methods in the 10x data set, and is close in performance to Seurat in the Dropseq and CEL-seq2 data sets. As illustrated in figure 2(c), performance is similar in the five cell mixture data sets, with the negative binomial model providing a clear advantage over the multinomial in all cases, and outperforming SC3 and Seurat on 10x data, whilst performing similarly to Seuart and SC3 on the CEL-seq2 data sets.

Another benchmarking sample in Tian et al.[19] is constructed as a mixture of bulk RNA from three different cell lines, combined in different proportions to produce seven different pseudo cells. Results comparing the performance of the methods we consider are shown in figure 2(b). Compared to the three and five cell mixtures, the performance of the multinomial model on this data is improved, but our zero-inflated negative binomial model still outperforms it. We hypothesize that the improvement in the multinomial model is due to the artificial construction of the samples used to produce the data. Specifically by producing pseudo cells through effectively summing bulk RNA quantities from different cell lines, the data are no longer counts from single cells, but rather a weighted sum of counts from multiple cell types. Performing model selection on this data set using the BIC in supplementary table 1 it can be seen that for this data the multinomial model is generally preferred, and this could explain the increased performance of the multinomial model on this data. However we would not generally expect such data to be considered in standard scRNA-seq analysis.

We also consider the number of clusters learnt by each model, with the results

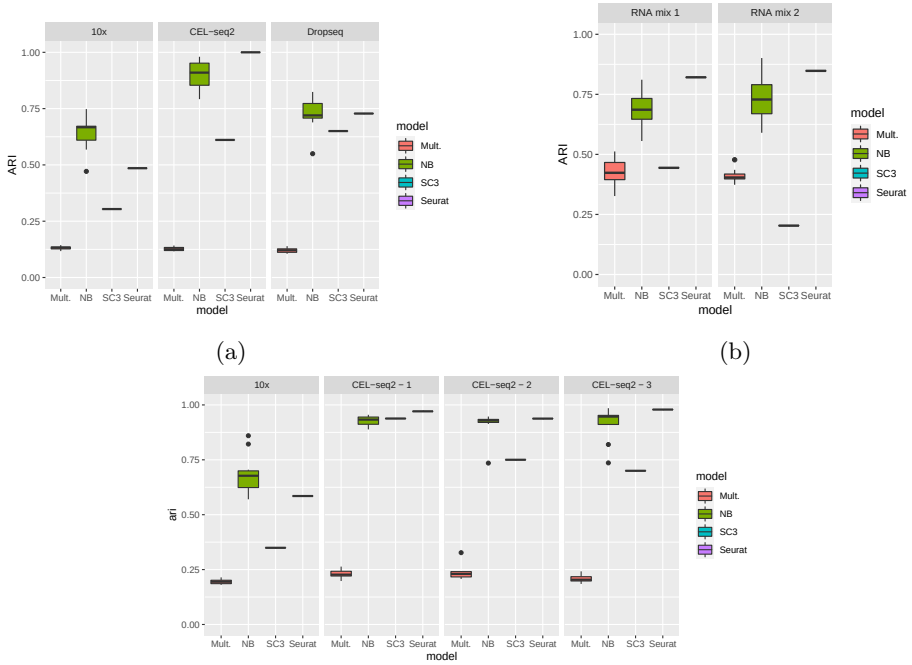


Figure 2: Performance measured by Adjusted Rand Index for 10 runs of each clustering model, across multiple data sets – (a) a mixture of 3 cell types across multiple experimental platforms, (b) a mixture of RNA in 7 different proportions from bulk cell populations for 3 different cell types (c) a mixture of 5 cell types across several experimental platforms.

illustrated in figure 3. It can be seen that across all three of the benchmarking data sets, our zero-inflated negative binomial approach is closer to identifying the correct number of clusters than the multinomial model, which significantly overestimates the number of clusters. The multinomial model appears to maintain the same number of clusters used in the K-means algorithm to initialise the sampler, and while in experiments we found that simply initialising the cluster membership at random does allow the multinomial model to infer fewer clusters, the final number still overestimated the number of clusters in the data by a large margin. For the three cell and five cell mixtures sequenced on the 10x platform, it can be seen that in terms of identifying the number of clusters, our method outperforms SC3, while performing similarly to SC3 and Seurat on CEL-seq2 and Dropseq data.

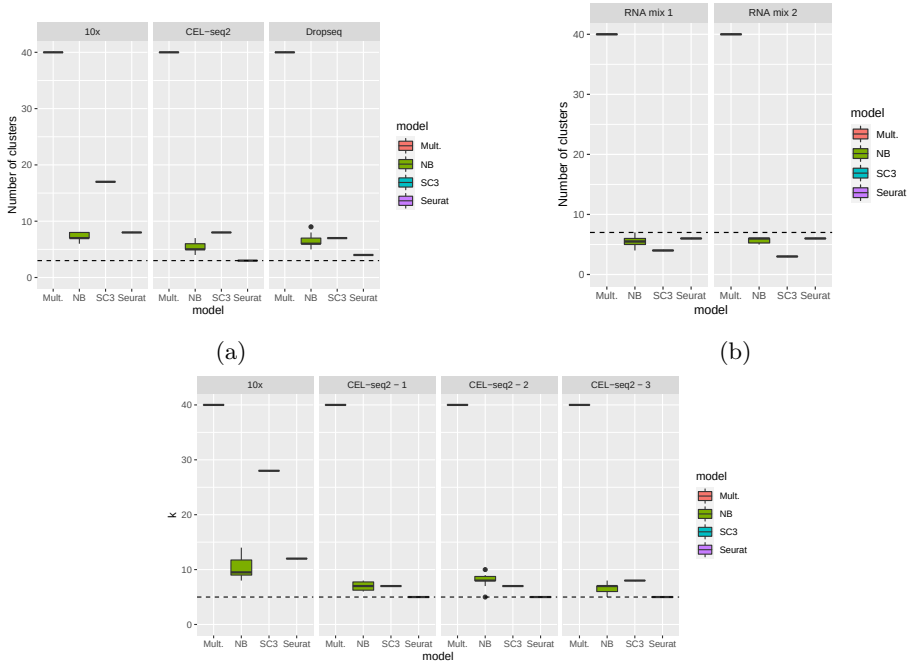


Figure 3: Numbers of clusters inferred (horizontal lines indicate true cluster number) for 10 runs of each clustering model across multiple data sets – (a) a mixture of 3 cell types across multiple experimental platforms, (b) a mixture of RNA in 7 different proportions from bulk cell populations for 3 different cell types (c) a mixture of 5 cell types across several experimental platforms.

3.2.2 Mouse cortex and hippocampus single cell data

To illustrate the results that we can produce using directly interpretable models, our method was applied to the single cell RNA-Seq data set of Zeisel et al.[20], that provides data for 3005 cells from the mouse cortex and hippocampus. In Zeisel et al.[20] a clustering produced using the BacksSPIN algorithm, a bi-clustering based variant of the SPIN algorithm[28], is refined using heuristic approaches to produce the clustering shown in figure 4. Applying our method to the data produces 20 clusters, many of which are clearly aligned to the cell types identified in the data in Zeisel et al.[20]. In some cases our clustering subdivides the higher level cell type specific clusters identified by Zeisel et al., who also observed lower level subdivisions of cell types into further clusters. A benefit of our approach is that we can investigate the MAP estimates of the

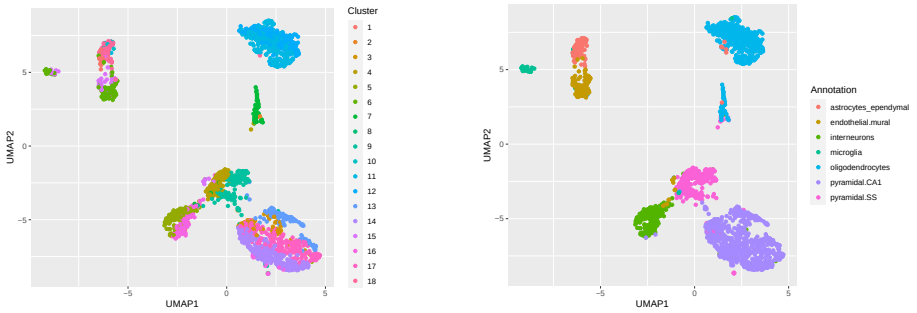
mean and dispersion parameters for each cluster, allowing us to easily identify key genes that are important in defining cluster membership. We can do so by comparing the mean expression of the gene in a cluster with the average expression across all clusters, and taking into account the mean and dispersion of the counts, identify genes whose expression levels largely differ from the across-cluster average.

We are also able to remove cells whose cluster membership cannot be identified with confidence. To do so we calculate the mean posterior probability of a cell belonging to the same cluster as the other cells in its MAP estimated cluster. Then we can filter out cells where this mean probability is below a certain threshold, here set to 0.5.

A UMAP[29] representation of the data combined with our clustering output is shown in figure 4. We focus on two specific clusters, 5 and 11, which are identified as interneurons and oligodendrocytes respectively in Zeisel et al.[20]. Considering transcripts where the MAP estimate of mean expression in a cluster is above the 95% confidence interval of the across cluster average expression distribution as overexpressed, our method identifies a set of overexpressed genes in both clusters. In cluster 5, upregulated genes include *Gad1*, *Gad2* and *Slc6a1*, known markers of interneurons. In cluster 11 on the other hand oligodendrocyte specific gene *Mog* and the gene *Mbp* that encodes a protein that is a major component of the myelin sheath, are discovered to be upregulated. This demonstrates that our approach can be used to identify genes that are indicative of cluster membership, that also have direct biological relevance to the underlying cell types.

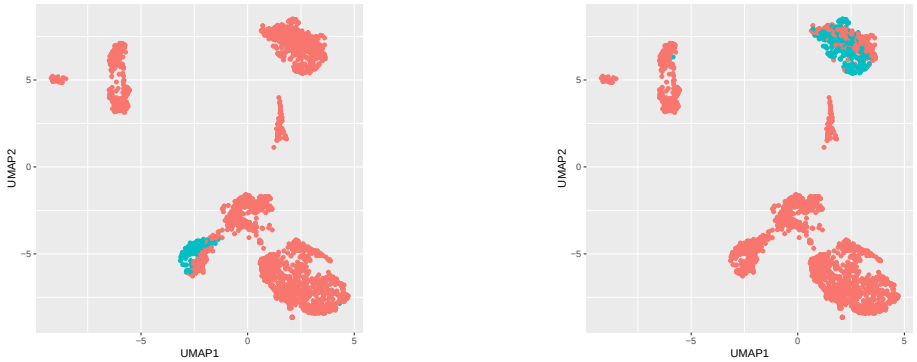
4 Discussion

In this paper we have introduced a novel Bayesian nonparametric zero-inflated negative binomial model for the clustering of scRNA-seq data, comprised of a mixture model at the count level for individual genes to allow for variation



(a) (b)

Figure 4: UMAP embeddings of cells from the mouse cortex and hippocampus data with (a) clusters generated by our zero-inflated negative binomial model, and (b) annotations from Zeisel et al. [20].



(a) (b)

Figure 5: UMAP embeddings of cells from the mouse cortex and hippocampus data[20] with clusters (a) 5 and (b) 11 identified by our method highlighted in blue.

in the degree of zero-inflation, as well as a second mixture at the cell level, and demonstrated that it performs as well as or better than other competing approaches when applied to benchmarking data.

A key benefit of our approach is that we can derive a statistical model of gene expression within each sub-population of cells identified. This allows us to easily identify genes that are over or under expressed in a sub-population of cells without post-hoc examination of the clustering. By applying a Bayesian model we are also able to quantify the uncertainty in cluster allocation for each cell. Furthremore we demonstrate that the zero-inflated negative binomial model we

propose outperforms the multinomial model of scRNA-seq counts, and through the use of statistical model selection show that for some genes a zero-inflated model is preferable to a negative binomial.

Applying the method to real world scRNA-seq data we are able to identify clusters of cells corresponding to cell types identified in previous work in the literature, and find cluster specific overexpressed genes that correspond to known cell type markers.

5 Data availability statement

The single cell benchmarking data from Tian et al.[19] are available from Gene Expression Omnibus `https://www.ncbi.nlm.nih.gov/geo/` with accession GSE118767, and from `https://github.com/LuyiTian/sc_mixology`. The mouse cortex and hippocampus data from Zeisel et al.[20] are available from GEO with accession GSE60361 and from `https://linnarssonlab.org/cortex/`. The code used to produce the results shown in the manuscript is available as a Snakemake pipeline[30] from `https://github.com/tt104/scmixture`. The Snakemake pipeline is also archived along with a Singularity container to run the code on Zenodo, DOI 10.5281/zenodo.5148814.

References

[1] Svensson V, Vento-Tormo R, Teichmann SA. Exponential Scaling of Single-Cell RNA-Seq in the Past Decade. Nature Protocols. 2018 Apr;13(4):599–604.

[2] Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC, et al. Integrative Single-Cell Analysis of Transcriptional and Epigenetic States in the Human Adult Brain. Nature Biotechnology. 2018 Jan;36(1):70–80.

[3] Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ,

et al. Single-Cell Transcriptomic Analysis of Alzheimer's Disease. Nature. 2019 Jun;570(7761):332–337.

[4] Stegle O, Teichmann SA, Marioni JC. Computational and Analytical Challenges in Single-Cell Transcriptomics. Nature Reviews Genetics. 2015 Mar;16(3):133–145.

[5] Szabo PA, Levitin HM, Miron M, Snyder ME, Senda T, Yuan J, et al. Single-Cell Transcriptomics of Human T Cells Reveals Tissue and Activation Signatures in Health and Disease. Nature Communications. 2019 Oct;10(1):4706.

[6] Liu Y, He S, Wang XL, Peng W, Chen QY, Chi DM, et al. Tumour Heterogeneity and Intercellular Networks of Nasopharyngeal Carcinoma at Single Cell Resolution. Nature Communications. 2021 Feb;12(1):741.

[7] Duan T, Pinto JP, Xie X. Parallel Clustering of Single Cell Transcriptomic Data with Split-Merge Sampling on Dirichlet Process Mixtures. Bioinformatics. 2019 Mar;35(6):953–961.

[8] Wu Q, Luo X. Nonparametric Bayesian Two-Level Clustering for Subject-Level Single-Cell Expression Data. arXiv:191208050 [stat]. 2021 Feb.

[9] Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells. Nature Biotechnology. 2014 Apr;32(4):381–386.

[10] Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: Consensus Clustering of Single-Cell RNA-Seq Data. Nature Methods. 2017 May;14(5):483–486.

[11] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019 Jun;177(7):1888–1902.e21.

[12] Zhang S, Li X, Lin Q, Wong KC. Review of Single-Cell RNA-Seq Data Clustering for Cell Type Identification and Characterization. arXiv:200101006 [cs, q-bio, stat]. 2020 Jan.

[13] Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep Generative Modeling for Single-Cell Transcriptomics. Nature Methods. 2018 Dec;15(12):1053–1058.

[14] Grønbech CH, Vording MF, Timshel P, Sønderby CK, Pers TH, Winther O. scVAE: Variational Auto-Encoders for Single-Cell Gene Expression Data. Bioinformatics. 2020 Aug;36(16):4415–4422.

[15] Svensson V, Gayoso A, Yosef N, Pachter L. Interpretable Factor Models of Single-Cell RNA-Seq via Variational Autoencoders. Bioinformatics. 2020 Jun;36(11):3418–3421.

[16] Levitin HM, Yuan J, Cheng YL, Ruiz FJ, Bush EC, Bruce JN, et al. De Novo Gene Signature Identification from Single-Cell RNA-Seq with Hierarchical Poisson Factorization. Molecular Systems Biology. 2019 Feb;15(2):e8557.

[17] Svensson V. Droplet scRNA-Seq Is Not Zero-Inflated. Nature Biotechnology. 2020 Feb;38(2):147–150.

[18] Clivio O, Lopez R, Regier J, Gayoso A, Jordan MI, Yosef N. Detecting Zero-Inflated Genes in Single-Cell Transcriptomics Data. bioRxiv. 2019 Nov:794875.

[19] Tian L, Dong X, Freytag S, Lê Cao KA, Su S, JalalAbadi A, et al. Benchmarking Single Cell RNA-Sequencing Analysis Pipelines Using Mixture Control Experiments. Nature Methods. 2019 Jun;16(6):479–487.

[20] Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, Manno GL, Juréus A, et al. Cell Types in the Mouse Cortex and Hippocampus Revealed by Single-Cell RNA-Seq. Science. 2015 Mar;347(6226):1138–1142.

[21] Lun ATL, McCarthy DJ, Marioni JC. A Step-by-Step Workflow for Low-Level Analysis of Single-Cell RNA-Seq Data with Bioconductor. F1000Research. 2016 Oct;5:2122.

[22] Liverani S, Hastie DI, Azizi L, Papathomas M, Richardson S. PReMiuM: An R Package for Profile Regression Mixture Models Using Dirichlet Processes. Journal of Statistical Software. 2015 Mar;64(7):1–30.

[23] Neal RM. Slice Sampling. The Annals of Statistics. 2003 Jun;31(3):705–767.

[24] Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, et al. CEL-Seq2: Sensitive Highly-Multiplexed Single-Cell RNA-Seq. Genome Biology. 2016 Apr;17(1):77.

[25] Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively Parallel Digital Transcriptional Profiling of Single Cells. Nature Communications. 2017 Jan;8(1):14049.

[26] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015 May;161(5):1202–1214.

[27] Hubert L, Arabie P. Comparing Partitions. Journal of Classification. 1985 Dec;2(1):193–218.

[28] Tsafrir D, Tsafrir I, Ein-Dor L, Zuk O, Notterman DA, Domany E. Sorting Points into Neighborhoods (SPIN): Data Analysis and Visualization by Ordering Distance Matrices. Bioinformatics. 2005 May;21(10):2301–2308.

[29] McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:180203426 [cs, stat]. 2020 Sep.

[30] Köster J, Rahmann S. Snakemake—a Scalable Bioinformatics Workflow Engine. Bioinformatics. 2012 Oct;28(19):2520–2522.

6 Supplementary Materials

Table 1: Bayesian Information Criterion scores for count data of individual clusters within data sets. Lower values indicate the model which is more likely to be the true model.

Multinomial BIC	Poisson BIC	Neg. Bin. ZI BIC	Neg. Bin. BIC	Cell type	Data set
$2.42264 \cdot 10^6$	$2.38944 \cdot 10^6$	$1.44142 \cdot 10^6$	$1.4372 \cdot 10^6$	HCC827	3 cell type 10x*
$2.88105 \cdot 10^6$	$2.79589 \cdot 10^6$	$1.64428 \cdot 10^6$	$1.63966 \cdot 10^6$	H1975	3 cell type 10x*
$2.73165 \cdot 10^6$	$2.70756 \cdot 10^6$	$1.5813 \cdot 10^6$	$1.57629 \cdot 10^6$	H2228	3 cell type 10x*
$7.80269 \cdot 10^5$	$7.94034 \cdot 10^5$	$5.83982 \cdot 10^5$	$5.80059 \cdot 10^5$	H1975	3 cell type CELseq2*
$5.89789 \cdot 10^5$	$5.95823 \cdot 10^5$	$4.25492 \cdot 10^5$	$4.21795 \cdot 10^5$	H2228	3 cell type CELseq2*
$6.17788 \cdot 10^5$	$6.60729 \cdot 10^5$	$4.39139 \cdot 10^5$	$4.35491 \cdot 10^5$	HCC827	3 cell type CELseq2*
$5.62865 \cdot 10^5$	$5.45388 \cdot 10^5$	$4.40203 \cdot 10^5$	$4.36755 \cdot 10^5$	H1975	3 cell type Dropseq*
$4.03465 \cdot 10^5$	$3.93161 \cdot 10^5$	$3.37921 \cdot 10^5$	$3.34598 \cdot 10^5$	HCC827	3 cell type Dropseq*
$3.93785 \cdot 10^5$	$3.82267 \cdot 10^5$	$3.26826 \cdot 10^5$	$3.23447 \cdot 10^5$	H2228	3 cell type Dropseq*
$3.22186 \cdot 10^5$	$3.26629 \cdot 10^5$	$2.45783 \cdot 10^5$	$2.42689 \cdot 10^5$	H838	5 cell type CELseq2 - 1*
$3.45298 \cdot 10^5$	$3.33441 \cdot 10^5$	$2.58841 \cdot 10^5$	$2.55728 \cdot 10^5$	H1975	5 cell type CELseq2 - 1*
$5.5902 \cdot 10^5$	$5.49083 \cdot 10^5$	$4.61282 \cdot 10^5$	$4.57408 \cdot 10^5$	A549	5 cell type CELseq2 - 1*
$3.30113 \cdot 10^5$	$3.35141 \cdot 10^5$	$2.42188 \cdot 10^5$	$2.39179 \cdot 10^5$	H2228	5 cell type CELseq2 - 1*
$2.91131 \cdot 10^5$	$3.112 \cdot 10^5$	$2.15471 \cdot 10^5$	$2.12676 \cdot 10^5$	HCC827	5 cell type CELseq2 - 1*
$2.3732 \cdot 10^5$	$2.35602 \cdot 10^5$	$1.96629 \cdot 10^5$	$1.93608 \cdot 10^5$	H2228	5 cell type CELseq2 - 2*
$2.57411 \cdot 10^5$	$2.56777 \cdot 10^5$	$2.14415 \cdot 10^5$	$2.11429 \cdot 10^5$	H1975	5 cell type CELseq2 - 2*
$5.16643 \cdot 10^5$	$5.39083 \cdot 10^5$	$4.46702 \cdot 10^5$	$4.42551 \cdot 10^5$	A549	5 cell type CELseq2 - 2*
$3.65866 \cdot 10^5$	$4.01282 \cdot 10^5$	$3.05143 \cdot 10^5$	$3.01695 \cdot 10^5$	H838	5 cell type CELseq2 - 2*
$2.54126 \cdot 10^5$	$2.73651 \cdot 10^5$	$2.00768 \cdot 10^5$	$1.97736 \cdot 10^5$	HCC827	5 cell type CELseq2 - 2*
$2.16319 \cdot 10^5$	$2.27348 \cdot 10^5$	$1.65913 \cdot 10^5$	$1.63055 \cdot 10^5$	H1975	5 cell type CELseq2 - 3*
$2.56021 \cdot 10^5$	$2.73285 \cdot 10^5$	$1.9104 \cdot 10^5$	$1.87885 \cdot 10^5$	H2228	5 cell type CELseq2 - 3*
$2.40969 \cdot 10^5$	$2.71051 \cdot 10^5$	$1.76488 \cdot 10^5$	$1.73558 \cdot 10^5$	HCC827	5 cell type CELseq2 - 3*
$3.96211 \cdot 10^5$	$4.18614 \cdot 10^5$	$3.24743 \cdot 10^5$	$3.21082 \cdot 10^5$	H838	5 cell type CELseq2 - 3*
$5.74114 \cdot 10^5$	$5.99038 \cdot 10^5$	$4.82863 \cdot 10^5$	$4.78676 \cdot 10^5$	A549	5 cell type CELseq2 - 3*
$3.48174 \cdot 10^6$	$3.47983 \cdot 10^6$	$2.89019 \cdot 10^6$	$2.88471 \cdot 10^6$	HCC827	5 cell type 10x
$5.1514 \cdot 10^6$	$5.17234 \cdot 10^6$	$4.24617 \cdot 10^6$	$4.24059 \cdot 10^6$	H838	5 cell type 10x
$2.95713 \cdot 10^6$	$2.90693 \cdot 10^6$	$2.18351 \cdot 10^6$	$2.17833 \cdot 10^6$	H1975	5 cell type 10x
$4.38239 \cdot 10^6$	$4.32627 \cdot 10^6$	$3.51452 \cdot 10^6$	$3.50859 \cdot 10^6$	H2228	5 cell type 10x
$6.26556 \cdot 10^6$	$6.24203 \cdot 10^6$	$5.41783 \cdot 10^6$	$5.41166 \cdot 10^6$	A549	5 cell type 10x
$1.78683 \cdot 10^5$	$1.81709 \cdot 10^5$	$1.82782 \cdot 10^5$	$1.79576 \cdot 10^5$	Mixture 0.68-0.16-0.16	RNA mix 1*
$3.71382 \cdot 10^5$	$3.75616 \cdot 10^5$	$3.75953 \cdot 10^5$	$3.7203 \cdot 10^5$	Mixture 0.33-0.33-0.33	RNA mix 1*
$2.02381 \cdot 10^5$	$2.05527 \cdot 10^5$	$2.06423 \cdot 10^5$	$2.03134 \cdot 10^5$	Mixture 0.16-0.68-0.16	RNA mix 1*
$2.00739 \cdot 10^5$	$2.04015 \cdot 10^5$	$2.05305 \cdot 10^5$	$2.02079 \cdot 10^5$	Mixture 1-0-0	RNA mix 1*
$1.91325 \cdot 10^5$	$1.94444 \cdot 10^5$	$1.96424 \cdot 10^5$	$1.93127 \cdot 10^5$	Mixture 0-1-0	RNA mix 1*
$1.92225 \cdot 10^5$	$1.95245 \cdot 10^5$	$1.96946 \cdot 10^5$	$1.93662 \cdot 10^5$	Mixture 0-0-1	RNA mix 1*
$1.68945 \cdot 10^5$	$1.71481 \cdot 10^5$	$1.72736 \cdot 10^5$	$1.69699 \cdot 10^5$	Mixture 0.16-0.16-0.68	RNA mix 1*
$1.99156 \cdot 10^5$	$2.00982 \cdot 10^5$	$1.95586 \cdot 10^5$	$1.92483 \cdot 10^5$	Mixture 0.68-0.16-0.16	RNA mix 2*
$3.731 \cdot 10^5$	$3.75458 \cdot 10^5$	$3.71666 \cdot 10^5$	$3.67829 \cdot 10^5$	Mixture 0.33-0.33-0.33	RNA mix 2*
$2.12002 \cdot 10^5$	$2.13354 \cdot 10^5$	$2.10773 \cdot 10^5$	$2.07674 \cdot 10^5$	Mixture 0.16-0.68-0.16	RNA mix 2*
$1.76565 \cdot 10^5$	$1.7832 \cdot 10^5$	$1.77576 \cdot 10^5$	$1.74494 \cdot 10^5$	Mixture 1-0-0	RNA mix 2*
$9.85492 \cdot 10^4$	$9.96553 \cdot 10^4$	$1.00335 \cdot 10^5$	$9.76862 \cdot 10^4$	Mixture 0-1-0	RNA mix 2*
$2.07126 \cdot 10^5$	$2.09322 \cdot 10^5$	$2.07834 \cdot 10^5$	$2.0452 \cdot 10^5$	Mixture 0-0-1	RNA mix 2*
$1.78978 \cdot 10^5$	$1.78982 \cdot 10^5$	$1.76544 \cdot 10^5$	$1.73579 \cdot 10^5$	Mixture 0.16-0.16-0.68	RNA mix 2*
$2.54561 \cdot 10^6$	$2.17056 \cdot 10^6$	$1.46266 \cdot 10^6$	$1.4585 \cdot 10^6$	interneurons	Mouse cortex and hippocampus†
$2.63688 \cdot 10^6$	$2.4405 \cdot 10^6$	$1.90501 \cdot 10^6$	$1.90081 \cdot 10^6$	pyramidal SS	Mouse cortex and hippocampus†
$5.99123 \cdot 10^6$	$5.82026 \cdot 10^6$	$4.68802 \cdot 10^6$	$4.6839 \cdot 10^6$	pyramidal CA1	Mouse cortex and hippocampus†
$5.42535 \cdot 10^6$	$5.00106 \cdot 10^6$	$3.52986 \cdot 10^6$	$3.52589 \cdot 10^6$	oligodendrocytes	Mouse cortex and hippocampus†
$5.68116 \cdot 10^5$	$4.73877 \cdot 10^5$	$3.62225 \cdot 10^5$	$3.5894 \cdot 10^5$	microglia	Mouse cortex and hippocampus†
$1.32336 \cdot 10^6$	$1.10568 \cdot 10^6$	$8.3758 \cdot 10^5$	$8.33743 \cdot 10^5$	endothelial mural	Mouse cortex and hippocampus†
$1.35327 \cdot 10^6$	$1.15411 \cdot 10^6$	$8.36217 \cdot 10^5$	$8.31806 \cdot 10^5$	astrocytes ependymal	Mouse cortex and hippocampus†

* Tian et al.[19]. † Zeisel et al.[20]