Likelihood-based cell type classification using single-cell RNA-sequencing data

By

Chang Yu

B.S. University of Massachusetts Amherst, 2019

Submitted in partial fulfillment of the requirements for the Master of Science in the Department of Biostatistics at Brown University School of Public Health

PROVIDENCE, RHODE ISLAND

MAY, 2021

This thesis by Chang Yu is accepted in its present form by the Department of Biostatistics as satisfying the thesis requirements for the degree of Master of Science

Date: _____ _____

Zhijin Wu, Ph.D., Advisor

Date: _____ _____

Roberta De Vito, Ph.D., Reader

Date: _____ _____

Stavroula Chrysanthopoulou, Ph.D., Program Director

Approved by the Graduate Council

Date: _____ _____

Andrew G. Campbell, Ph.D., Dean of Graduate School

Table of Contents

List of Tables

List of Figures

1 Introduction

Single-cell RNA sequencing (scRNA-seq) quantifies the whole transcriptome on a cellular-level. The high-resolution genome-scale data enables the exploration of tissue heterogeneity in biological samples with interesting applications, such as proposing rare cell types [1], studying expression variation across cell populations [2], and cell fate-mapping and lineage-tracing [3].

One common application in scRNA-seq is cell type identification. When cells are sequenced in a new experiment, we hope to identify the cell type of each cell based on its expression profile. Current annotation methods are either clustering-based or reference-based. The clustering-based approach first aggregates the unlabeled cells into groups according to their expression profile in an unsupervised manner. Next, differentially expressed genes between clusters are compared against the known marker genes to define a cell type for each cluster. On the other hand, the reference-based methods use cells with known cell type labels as reference. These methods then compare each new cell with the reference to propose a label. It has been shown that the supervised approach can control the level of granularity of cell types proposed. Furthermore, the number of cell types identified by the supervised approach does not grow with an increasing dataset, in contrast to the clustering-based approaches [4].

A number of supervised cell classifiers have been developed to identify cell types in scRNA-seq data. Methods including CellAssign [5] and Garnett [6] utilize marker genes that uniquely identify cell types. With no additional training dataset required, they rely heavily on well curated gene lists. An alternative is classifiers based on machine learning algorithms; for example, CaSTLe [7] trains XGBoost classification models, while scVI [8] uses deep neural networks. However, such complex models may lead to over-fitting. Some other methods [9–11] classify a new cell based on the correlation between its expression and the reference cells.

For example, scmap first summarizes each cell type in the training scRNAseq data as a centroid by finding the median expressions. Then, for each new cell, the cosine similarity, Pearson, and Spearman correlations are calculated between it's selected features and each centroid. A cell type is proposed if at least two of the three previously described measures agree and their maximum passes a given threshold; otherwise, the cell will remain unassigned [9]. CHETAH uses the averaged gene expressions for each cell type to build a hierarchical classification tree based on the Spearman

correlation and the average linkage. Each new cell is sent from the root node down the classification tree, with the path and ending position determined by calculated and expected correlations. With a given threshold, a cell is either assigned into a cell type, an intermediate node or remains unassigned [10]. SingleR allows using either scRNAseq data or bulk transcriptomic data on purified cells (microarray or RNAseq) as training data. Without further summarizing the reference data, a Spearman coefficient is calculated for a new cell with each reference sample and several cell types are proposed according to the summarized similarity scores. In the fine-tuning step, correlation analysis is repeated with the reduced cell types and variable genes between them. One cell type is eliminated each step until only one is left as the proposed cell type [11]. The correlation-based methods may be sensitive to the choice of genes to include in the model.

Instead of correlation, Grabski and Irizarry's method is based on a probabilistic model, using sorted single-cell data as the training set. Assuming the gene counts conditioned on cell types follow a Poisson distribution where the rate parameter comes from a mixture of three distributions, each representing this gene in an on (expressed), off-low, or off-high state. For each gene, five parameters describing the mixed distributions, which are shared across cell types, are pre-trained with large public datasets. Two additional parameters describing the gene and cell type specific mixing proportions are estimated with the training data [4].

When there is a hierarchical structure in the cell types given in the reference dataset, while the methods mentioned above are able to correctly identify the major cell types, it can still be challenging to distinguish cells that belong to different subtypes. Moreover, for the correlation-based and machine learning-based methods, it would be difficult to further interpret why a cell is assigned with a certain cell type. Furthermore, because different sequencing platforms may be associated with different characteristics, it is unclear whether the expression signatures identified from one platform can be directly used to classify cells sequenced by another platform.

For example, Drop-seq and 10x Genomics are often used to query large populations of cells, with a low coverage for each cell. Some scRNA-seq experiments use unique molecular identifiers (UMIs), which allow us to reduce multiple counts of RNA reads that originate from the same transcript to a single count, while avoiding amplification bias. In contrast, experiments that do not use UMIs would count all copies and, therefore, yield higher counts that are not directly comparable. Ideally, we

would like a classification tool that is robust to the choice of sequencing method, such that cell typing can be performed even when we encounter data obtained from a different sequencing platform.

In this study, I present a likelihood-based approach to cell type classification with a focus on transferable classification between sequencing platforms. The models assume the gene expressions follow some gene and cell type specific parametric distributions, where the parameters can be estimated using cells from training data and their cell type labels. For a new cell, the likelihood of a cell coming from each cell type is calculated, and the most likely cell type will yield the predicted label.

The rest of the thesis is organized as follows: in *Section 2* we describe the datasets used along with the methodology, *Section 3* presents results on within and cross platform model fitting and prediction and compare with alternatives. Finally, in *Section 4* we discuss future directions and limitations to this study.

2 Data and Methodology

2.1 Data

Both datasets used in this study were obtained from the Allen Brain Atlas Cell Types Database [12]. The first dataset consists of 76,533 nuclei derived from the primary motor cortex (M1C or M1) from two postmortem brains. The nuclei were sequenced with 10x Genomics Chromium Single Cell 3' v3 RNA-sequencing platform and were referred to as the "10X" dataset for the rest of the study. The second dataset consists of 49,417 nuclei derived from multiple origins from three postmortem brains. The cells are sampled from regions including middle temporal gyrus (MTG), anterior cingulate cortex (ACC; also known as the ventral division of medial prefrontal cortex, A24), primary visual cortex (V1C), primary motor cortex (M1C), primary somatosensory cortex (S1C), and primary auditory cortex (A1C). The nuclei were sequenced with SMART-Seq v4 platform and were referred to as the "SMART-seq" dataset for the rest of the study.

Both datasets were processed and clustered with the *scrattch.hicat* package following the same protocol [12]. Clusters derived and corresponding inferred cell types could be obtained from the Allen Brain Atlas Cell Types Database as cell metadata. For all the analysis in this study, the cell type labels provided are considered to be the true cell type for each cell. Furthermore, all comparisons between the two datasets done in the study are based on the assumption that the cell type definition is consistent across sampled regions and donors.

A total of 127 cell types were identified from the 10X dataset, consisting of 72 GABAergic neuron types, 44 Glutamatergic neuron types, and 11 non-neuronal cell types. A total of 120 cell types were identified from the SMART-seq dataset, consisting of 54 GABAergic neuron types, 56 Glutamatergic neuron types, and 10 non-neuronal cell types. Because detailed cell type labels for neuronal cells provided are rather ambiguous, I decide to carry on with only non-neuronal cells from each dataset. This resulted in a reasonable number of cells and cell types to work with. Composition of non-neuronal cell types for 10X and SMART-seq datasets is shown in *Table 1*. One thing to note is that three astrocyte subtypes were identified in each dataset, however there is no evidence showing one-to-one correspondence of the subtypes between datasets (data not shown). Similarly, four oligodendrocyte subtypes were identified from 10X, while two were identified from SMART-seq, and they showed no one-to-one correspondence. Thus, in further analysis, if needed, the subtypes

Table 1: Count and composition of non-neuronal cell types for 10X and SMART-seq datasets

	Full Label	Astro1	Astro2	Astro3	Endo	Micro	Oligo1	Oligo2	Oligo3	Oligo4	OPC	VLMC	Total
10X	# Cells	119	377	72	64	108	2300	101	535	6	283	40	4005
	General Label	Astro			Endo	Micro	Oligo				OPC	VLMC	Total
	# Cells	568			64	108	2942				283	40	4005

	Full Label	AstroA	AstroB	AstroC	Endo	Micro	OligoA	OligoB	OPC	Peri	VLMC	Total
SMARTseq	# Cells	966	75	146	70	750	254	1676	773	32	11	4753
	General Label	Astro			Endo	Micro	Oligo		OPC	Peri	VLMC	Total
	# Cells	1187			70	750	1930		773	32	11	4753

	Full Label	Astro1	Astro2	Astro3	Endo	Micro	Oligo1	Oligo2	Oligo3	Oligo4	OPC	VLMC
10X	% Cells	2.97%	9.41%	1.80%	1.60%	2.70%	57.43%	2.52%	13.36%	0.15%	7.07%	1.00%
	General Label	Astro			Endo	Micro	Oligo				OPC	VLMC
	% Cells	14.18%			1.60%	2.70%	73.46%				7.07%	1.00%

	Full Label	AstroA	AstroB	AstroC	Endo	Micro	OligoA	OligoB	OPC	Peri	VLMC
SMARTseq	% Cells	20.32%	1.58%	3.07%	1.47%	15.78%	5.34%	35.26%	16.26%	0.67%	0.23%
	General Label	Astro			Endo	Micro	Oligo		OPC	Peri	VLMC
	% Cells	24.97%			1.47%	15.78%	40.61%		16.26%	0.67%	0.23%

Abbreviations: Astro - astrocyte, Endo - endothelial cells, Micro - microglia, Oligo - oligodendrocytes, OPC - oligodendrocyte precursor cells, Peri - pericytes, and VLMC - vascular and leptomeningeal cells.

can be combined to provide a more general cell type information. Pericyte is a cell type only present in the SMART-seq dataset, however they only make up a small proportion of the dataset, and thus, were removed when performing cross dataset model fitting and evaluation.

2.2 Normalization

From the raw counts, I perform Counts per Million (CPM) normalization. All values in the CPM matrix were added a value of one, followed by a log_2 transformation. The resulting normalized matrices were referred to as $log_2(cpm+1)$ matrices. As a result, the genes having zero-detection still have value zero in the $log_2(cpm+1)$ matrix.

2.3 Marker Gene Selection

Marker genes are selected in a hierarchical manner, using the $log_2(cpm+1)$ matrices. All differential expression analysis is done with the *limma* package and fit with empirical Bayes. First across general cell types, for each gene, a linear model is fitted and a μ_i is estimated for each cell type i, where $\mu = \frac{\sum_{i=1}^{K} \mu_i}{K}$. F-test is performed for $H_0 : \mu_i = \mu, \forall i$, and genes are sorted by the p-value rejecting the

null. The top ten genes that satisfy a p-value adjusted with Benjamini & Hochberg method [13], such that the p-value $< .05$ and $sign(\mu_i - \mu) = -sign(\mu_j - \mu), \forall j \neq i$, are chosen as marker genes for cell type i. The same procedure is repeated for choosing markers genes to distinguish sub-cell types. The union of the lists are used as marker genes to be included in the model.

2.4 Likelihood Models

The gene expression is modeled with some parametric distributions, with gene g and cell type k specific parameters $\theta_{\mathbf{gk}}$ which are vectors of length p. Assuming independence among all genes G, and denoting the gene expression of a cell $\mathbf{y}$, the joint log likelihood of such cell belonged to cell type k is expressed as the summation of log likelihood for each gene, such that:

$$l(\mathbf{y}|\theta_{.\mathbf{k}}) = \sum_{g=1}^{G} l(y_g|\theta_{\mathbf{gk}}) \tag{1}$$

The parameters $\theta_{\mathbf{gk}}$ are estimated in the K cell types given in the training data. In order to prevent over-fitting for each gene, similar cell types are collapsed to share parameters. This reduces the number of parameters from pK to a smaller set.

The most likely cell type is given by:

$$\arg \max_{k=1,\ldots,K} l(\mathbf{y}|\theta_{.\mathbf{k}}) \tag{2}$$

2.4.1 Zero-Inflated Poisson Model

Model Setup

Assuming gene expression count Y_{gkj} where g is gene index, k and j together point to a specific cell in cell type k, where Y_{gkj} follows a zero-inflated Poisson (zi-Poisson) distribution with the following probability mass function:

$$f_{zi-Poisson}(Y_{gkj} = y|\pi_{gk}, \lambda_{gk}, L_{kj}) = \pi_{gk}\delta_0(y) + (1 - \pi_{gk}) f_{Poisson}(y|\lambda_{gk}L_{kj})$$

where π_{gk} is the amount of zero inflation, $\delta_0(y)$ is the indicator of $y = 0$; $f_{Poisson}(.)$ is the probability mass function of a Poisson distribution, and L_{kj} is the library size.

Parameter Estimation

Using marginal probability of $Y_{gkj} = 0$ and expectation of Y_{gkj} to set up the following two equations:

$$P(Y_{gkj} = 0) = \pi_{gk} + (1 - \pi_{gk}) exp\{-\lambda_{gk} L_{kj}\}$$

$$E(Y_{gkj}) = (1 - \pi_{gk}) \lambda_{gk} L_{kj}$$

Given that n_k cells in training data were from cell type k, let $\bar{Y}_{gk} = \frac{1}{n_k} \sum_{j=1}^{n_k} Y_{gkj}$. A method of moment estimator for π_{gk} is derived by solving:

$$\frac{1}{n_k} \sum_{j=1}^{n_k} \delta_0(Y_{gkj}) = \pi_{gk} + (1 - \pi_{gk}) exp\{-\frac{\bar{Y}_{gk}}{1 - \pi_{gk}}\} \tag{3}$$

within constraints $[0, 1]$. Let solution to equation (3) be $\tilde{\pi}_{gk}$, then the estimator for π_{gk} is

$$\hat{\pi}_{gk} = \tilde{\pi}_{gk} \cdot I[0 \leq \tilde{\pi}_{gk} \leq 1] + 1 \cdot I[\sum_{j=1}^{n_k} \delta_0(Y_{gkj}) = n_k] + 0 \cdot I[\frac{1}{n_k} \sum_{j=1}^{n_k} \delta_0(Y_{gkj}) \leq exp\{-\bar{Y}_{gk}\}]$$

Considering transformation of data $Z_{gkj} = \frac{Y_{gkj}}{L_{kj}}$ and $\bar{Z}_{gk} = \frac{1}{n_k} \sum_{j=1}^{n_k} Z_{gkj}$, then $E(Z_{gkj}) = \lambda_{gk}$. With previously estimated $\hat{\pi}_{gk}$, the method of moment estimator for λ_{gk} is

$$\hat{\lambda}_{gk} = \frac{\bar{Z}_{gk}}{1 - \hat{\pi}_{gk}}$$

The log likelihood of a cell being cell type k for gene g is calculated as:

$$l_{gk}(Y_{gkj} = y | \hat{\pi}_{gk}, \hat{\lambda}_{gk}) = log\{[\hat{\pi}_{gk} \delta_0(y) + (1 - \hat{\pi}_{gk}) f_{Poisson}(y | \hat{\lambda}_{gk} L_{kj})]\} \vee c$$

where c is the smallest finite log likelihood calculated for all genes and cell types for all the cells to be predicted, as a boundary to avoid introducing infinity. Joint log likelihood for all genes is calculated according to equation (1).

2.4.2 Normal Mixed with Point Mass at Zero Model

Model Setup

Assuming gene expression either in the form of $log_2(cpm+1)$ or count Y_{gkj} follows a normal distribution mixed with a discrete point mass at zero (may be referred to as Norm0 for short), with the following probability density function:

$$f_{Norm0}(Y_{gkj} = y | p_{0gk}, \mu_{gk}, \sigma^2_{gk}) = p_{0gk}\delta_0(y) + (1 - p_{0gk}) f_{Norm}(y | \mu_{gk}, \sigma^2_{gk})$$

where p_{0gk} is the probability of observing zero from the point mass, and $f_{Norm}(.)$ is the probability density function of a Gaussian distribution. Although gene expression can only be non-negative, the density of normal on the negative side is ignored here in this model.

Parameter Estimation

To ensure the model probability mass is not completely dominated by the discrete point mass at zero nor the Gaussian distribution, a $Beta(1,1)$ prior is used for p_{0gk} which results in the following estimator:

$$\hat{p}_{0gk} = \frac{1 + \sum_{j=1}^{n_k} \delta_0(Y_{gkj})}{1 + n_k}$$

The mean and variance of the normal distribution are estimated with conditional mean and sample variance as:

$$\hat{\mu}_{gk} = \frac{\sum_{j=1}^{n_k} Y_{gkj}}{\sum_{j=1}^{n_k} I[Y_{gkj} > 0]}$$

$$\hat{\sigma}^2_{gk} = \frac{\sum_{j=1}^{n_k} [Y_{gkj} - \hat{\mu}_{gk}]^2 I[Y_{gkj} > 0]}{-1 + \sum_{j=1}^{n_k} I[Y_{gkj} > 0]}$$

When modeling count data, Poisson counting error is expected, thus we set

$$\hat{\sigma}^2_{gk,bounded} = \hat{\sigma}^2_{gk} \wedge \hat{\mu}_{gk} \tag{4}$$

For cell types i that never observed any non-zero expression of a gene, their $\hat{\mu}_{gi}$ is imputed by mean of $\hat{\mu}_{gk}$ which have valid estimations. Similarly, for all $\hat{\sigma}^2_{gi}$ that can not be estimated they are imputed by the mean of all valid $\hat{\sigma}^2_{gk}$ across all genes and cell types.

Model Selection

When estimating a unique set of parameters for each cell type, it's important to note that each gene may not have a distinct distribution for every cell type (Figure 1 and 2). Thus, we may not need K sets of parameters to describe the cell type-specific expression. We collapse the number of parameters to fewer sets where the difference between cell types is small in order to avoid over-fitting and increase robustness. That is, for each gene, we choose a model between a null model (lack of differential expression in any cell type) and full model (each cell type has a unique θ_k) that may allow the gene to have 1 to K sets of unique parameters.

Specifically, for each gene, let $\hat{\mu}_{\mathbf{g}}$ be the estimated means for the normal part of the model, with i unique levels. The two cell types that have the closest means are combined, resulting in reduced cell type labels, which are then used to estimate $\hat{\mu}'_{\mathbf{g}}$. A test statistic is constructed as follows:

$$T = \frac{\sum_{k,j}[(Y_{gkj} - \hat{\mu}'_{gk})^2 - (Y_{gkj} - \hat{\mu}_{gk})^2]I(Y_{gkj} > 0)}{(\sum_{k,j} I(Y_{gkj} > 0) - i)\sum_{k,j}(Y_{gkj} - \hat{\mu}_{gk})^2 I(Y_{gkj} > 0)}$$

Under the null hypothesis, which indicates that two groups combined have the same mean, $T \sim F_{df}$ with degrees of freedom, $df = \sum_{k,j} I(Y_{gkj} > 0) - i$. When $P(F_{df} > T)$ is greater than a tuning parameter *p-value cutoff*, the two cell type labels are combined and $\hat{\mu}_{\mathbf{g}}$ is updated.

This process is repeated by iteratively combining the two groups with the closest estimated means until either H_0 is rejected or when all cell types are combined into one. Then the final reduced cell type labels are used to estimate $\hat{\mu}_{gk}$ and $\hat{\sigma}^2_{gk}$, in order to achieve model selection on the normal part.

2.5 Cross-Validation and Model Comparison

For within dataset model fitting, one third of cells from each cell type is used as the test set, while the other two thirds are used as training data to fit the models mentioned above. The random splits are repeated 10 times, and seeds are used for reproducibility.

To summarize training data fit and test set prediction results, accuracy is calculated with the true cell labels and the fitted or predicted labels.

2.6 Cross Platform Prediction

In order to learn from one dataset, denoted as A, and predict cell types of cells from another dataset, B, first we check if the marker genes identified in A still carried information in B. Specifically, marker genes are identified with dataset A and its full cell type labels. The training and testing datasets are then constructed by the expression of such genes in dataset B. This is done in both directions between 10X and SMART-seq datasets.

The norm0 models using $log_2(cpm+1)$ matrices as input, and the zi-Poisson models are examined. Marker genes to include in the model are proposed with dataset A. Parameters are estimated with the cells' expressions and the full cell labels from dataset A. With or without parameter correction, the estimated parameters are used to predict cell types of cells from dataset B with expression matrix B as input. Predicted cell type labels are collapsed to general cell types and compared with true labels of cells from dataset B. This is done in both directions between 10X and SMART-seq datasets.

For norm0 models, $\hat{p}_{0gk}$ are adjusted the following way: first, the marginal $\hat{\mathbf{p}}_{\mathbf{0,10X}}$ and $\hat{\mathbf{p}}_{\mathbf{0,SMART-seq}}$ for each dataset are found, where $\hat{p}_{0g} = \frac{\sum_{k,j} \delta_0(Y_{gkj})}{\sum_k n_k}$. $\hat{\mathbf{p}}_{\mathbf{0,10X}}$ and $\hat{\mathbf{p}}_{\mathbf{0,SMART-seq}}$ are each sorted and then mapped from one to another as quantile normalization. For example, when parameters are originally estimated with 10X data, each $\hat{p}_{0gk}$ is then corrected using mapping from $\hat{\mathbf{p}}_{\mathbf{0,10X}}$ to $\hat{\mathbf{p}}_{\mathbf{0,SMART-seq}}$.

2.7 Other Models Compared

Currently available cell classification tools, including scmap-cluster (referred to as scmap for the following), Grabski & Irizarry (referred to as GnI)[4], CHETAH[10], and SingleR [11], are evaluated for within and between dataset fitting and prediction. GnI requires raw counts as input, while others use $log_2(cpm+1)$ as input. All methods are run with their default gene selection option, and all other default thresholds. CHETAH and scmap are also run with lowered threshold parameters, 0 for CHETAH and .1 for scmap, to force the model to leave less cells with unassigned type.

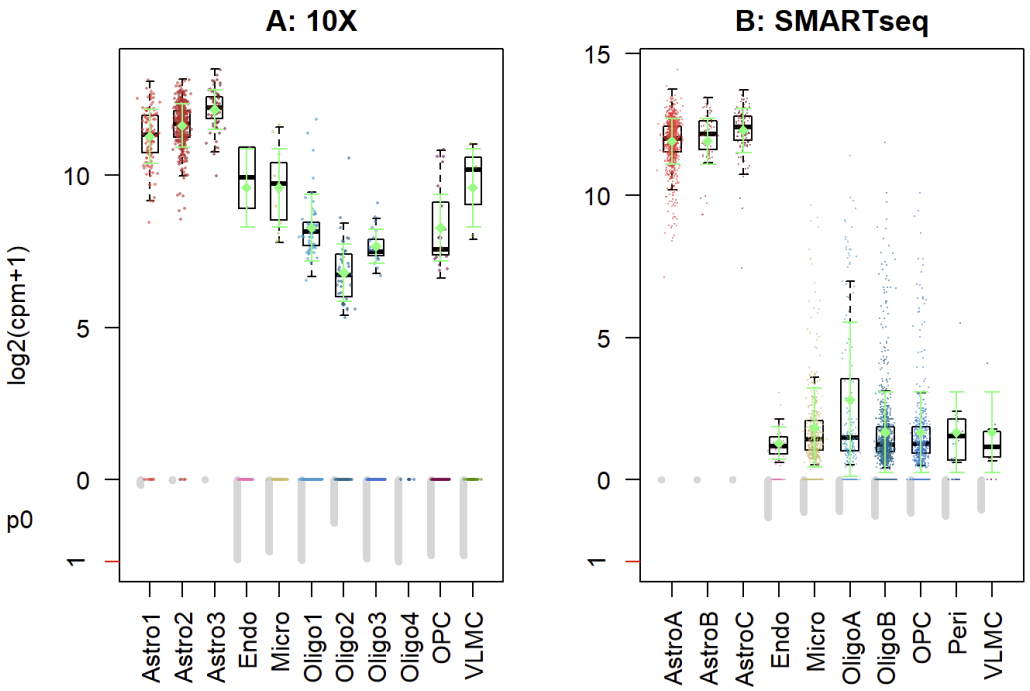

Figure 1: Expression distributions and estimated parameters for gene *ADGRV1* in 10X (A) and SMARTseq (B). Each dot represents a cell and the box shows expression level in $log_2(cpm+1)$ when detected in each cell type. The green dot and whiskers together show the $\hat{\mu} \pm \hat{\sigma}^2$ estimated with model selection using p-value cutoff .05. The gray bar shows the proportion of cells in each cell type that do not detect this gene.

3 Results

An example of distributions for gene *ADGRV1*'s expression across cell type is shown in *Figure 1*. Since *ADGRV1* has been shown to be a marker gene expressed by astrocytes [14], as expected, it was observed to have a high expression level in all three subtypes of astrocytes in both the 10X (*Figure 1A*) and the SMARTseq (*Figure 1B*) dataset. In 10X, *ADGRV1* was detected in about half of the Oligo2 cells, with a lower expression level when detected. For other cell types in the 10X dataset, *ADGRV1* was rarely detected (*Figure 1A*). However in the SMARTseq dataset, *ADGRV1* had non-zero expression for more than half of the cells for every cell type, while having lower expression levels in the non-Astro cell types (*Figure 1B*).

When the gene is detected, the expression levels are not distinct across all cell types. For example, in SMARTseq, the distributions of non-zero expressions of *ADGRV1* were similar in Micro, OligoB, OPC, Peri, and VLMC cells. As a result, when estimating the μ and σ^2 parameters for norm0 models

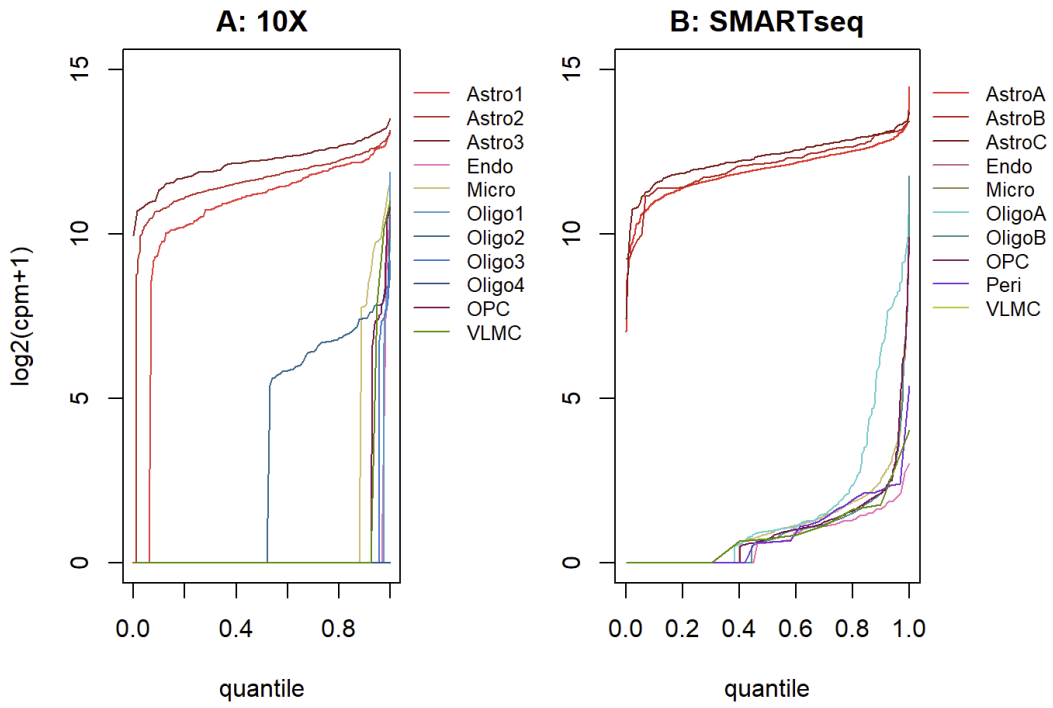

Figure 2: Quantiles of gene *ADGRV1*'s expression for each cell type in 10X (A) and SMARTseq (B).

with model selection, they were reduced to share the same $\hat{\mu}$ and $\hat{\sigma}^2$. Additionally, comparing the quantiles, OligoA had heavier upper tail, and Endo had a lighter upper tail, thus they were not reduced to other non-Astro cell types and had distinct parameter estimations (*Figure 2B*). Another example was that although *ADGRV1* had higher expression level in Astro cells in both datasets, its expression level was distinct in all three Astro subtypes in 10X (*Figures 1A and 2A*), but similar between AstroB and AstroC in SMARTseq (*Figures 1B and 2B*).

Furthermore, as expected, due to utilization of UMIs, the gene was not detected in more cells from the 10X dataset compared to the SMARTseq dataset (*Figures 1 and 2*). The skewed QQ-plot of p_0 calculated marginally across all cell types in each dataset also demonstrated higher detection rate in the SMARTseq dataset (*Figure 3*).

3.1 Within 10X Dataset

The zero-inflated Poisson (zi-Poisson) model and different versions of the normal mixed with point mass at zero (norm0) model were compared, for fitting and predicting cells within the 10X dataset (*Figure 4*). Overall, the norm0 models performs better than the zi-Poisson model, and the

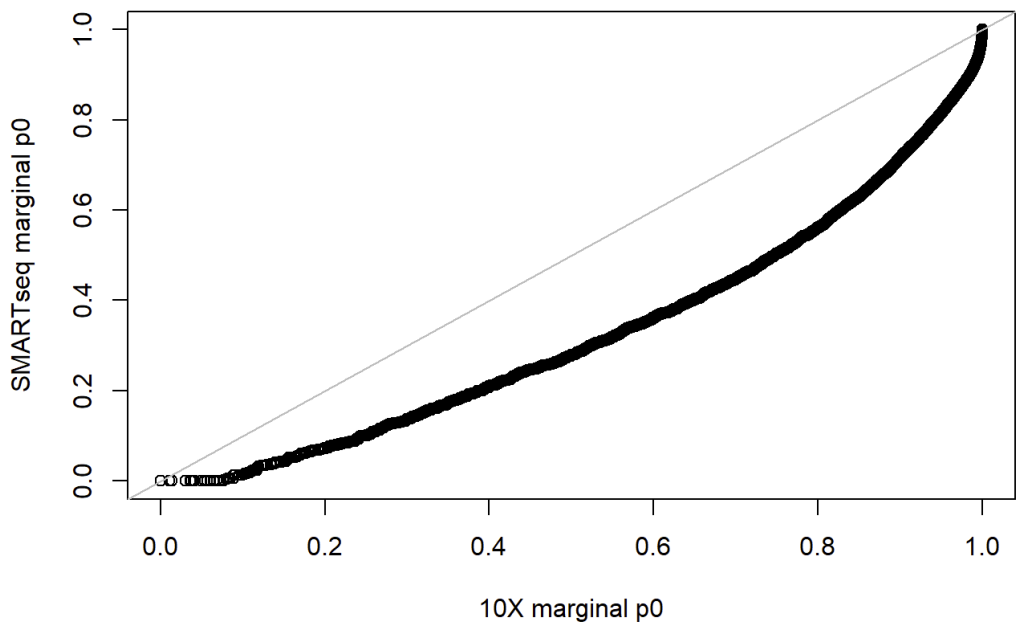

Figure 3: QQ-plot between marginally estimated p_0 in 10X and SMARTSeq. Straight line shows the identity line.

norm0 models using $log_2(cpm+1)$ matrices as input are able to fit better and make more accurate predictions. Moreover, including model selection for the $log_2(cpm+1)$ norm0 model (with p-value cutoff=$1e-4$) reduces over-fitting and results in more accurate predictions than without model selection (with p-value cutoff=1).

The zi-Poisson model and the best performing version of norm0 models, which used $log_2(cpm+1)$ input and model selection, have similar or better performances than the currently available tools (Figure 5). Both of our models outperform scmap, whether using its default or a lowered threshold parameter, and Grabski and Irizarry's method. They also achieve similar prediction accuracies to the best performing models, CHETAH and SingleR (*Figure 5*).

3.2 Within SMARTseq Dataset

It is shown again that including model selection (with p-value cutoff=$1e-4$) results in improved prediction accuracies in the norm0 models. Such norm0 models also fit and predict better than the zi-Poisson model. Moreover, for SMARTseq data, the norm0 models perform better when using raw counts as input than using $log_2(cpm+1)$ data, however setting a lower bound for $\hat{\sigma}^2$ as in equation (4) is not necessary (*Figure 6*)

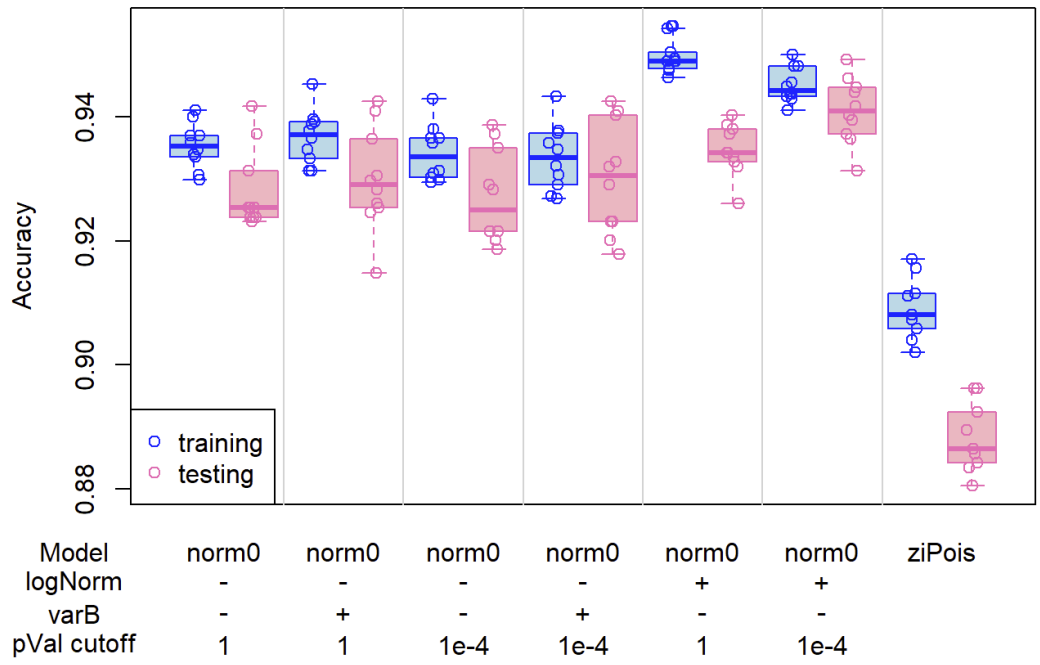

Figure 4: Accuracy for model fitting and prediction accuracy for likelihood models with 10X data. Points added on top of bar plots. The norm0 and zi-Poisson models are fitted with 127 marker genes as input. LogNorm indicates either used count matrix (-) as input or $log_2(cpm+1)$ matrix (+) as input. VarB indicates whether $\hat{\sigma}^2$ is adjusted as in equation (4) for the normal0 models. pVal cutoff indicates p-value cutoff used for model selection, where 1 represents no model selection. Blue bars and points indicates model fitting, and pink indicates testing accuracy.

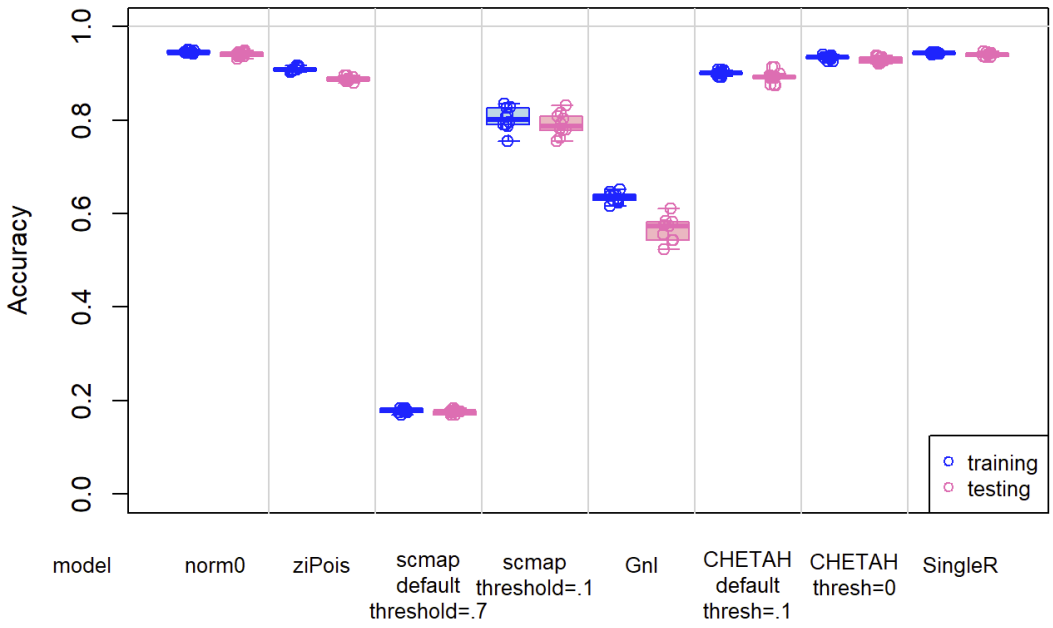

Figure 5: Comparing accuracy for model fitting and prediction accuracy between likelihood models and available methods with 10X data. Points added on top of bar plots. The norm0 model shown here uses $log_2(cpm+1)$ input and model selection with p-value cutoff $1e-4$. CHETAH and scmap are fitted twice using either their default or lowered threshold parameter. Blue bars and points indicates model fitting, and pink indicates testing accuracy.

The zi-Poisson model and the best performing version of norm0 models, which use raw counts input and model selection, have similar or better performances than currently available tools (*Figure 7*). Our models outperform scmap with its default threshold parameter, and had similar prediction accuracies to the best performing models, CHETAH and SingleR. Since the gene specific parameters for GnI were pre-trained with 10X data, as expected, the model fits the SMARTseq data poorly (*Figure 7*).

3.3 Cross Platfrom Prediction

3.3.1 From 10X to SMARTseq

When training norm0 or zi-Poisson models with 10X data and predict cell types of cells from SMARTseq data, both $\hat{p_0}$ and $\hat{\sigma}^2$ need to be adjusted in order to get good prediction results. However, including model selection is not necessary, although it is not harmful. Adjusting the $\hat{\sigma}^2$ alone, the norm0 models perform similar to the zi-Poisson model (*Figure 8*).

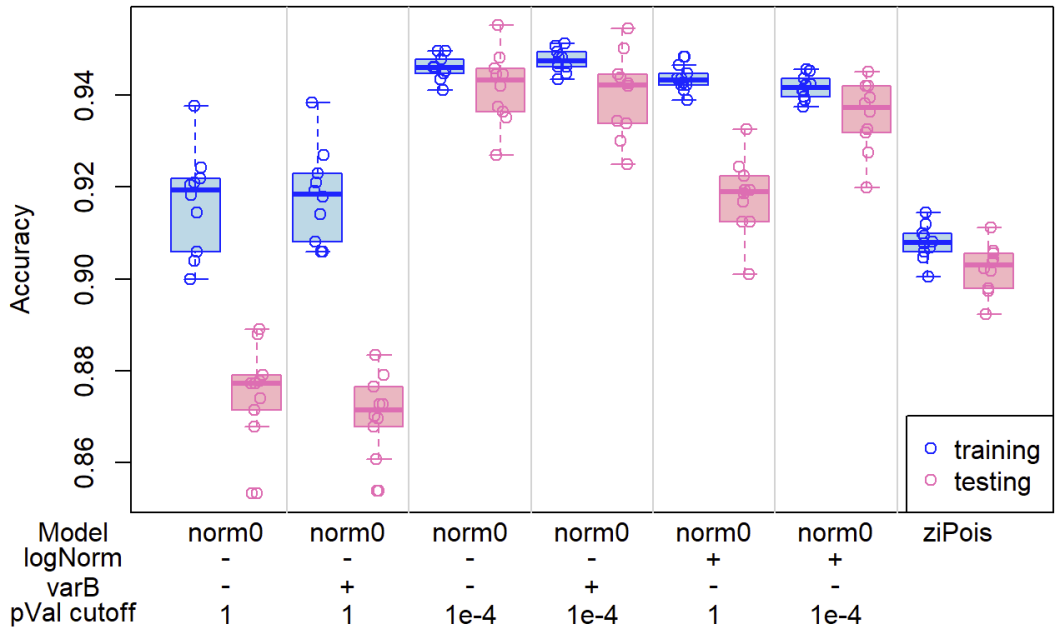

Figure 6: Accuracy for model fitting and prediction accuracy for likelihood models with SMARTseq data. Points added on top of bar plots. The norm0 and zi-Poisson models are fitted with 127 marker genes as input. LogNorm indicates either used count matrix (-) as input or $log_2(cpm+1)$ matrix (+) as input. VarB indicates whether $\hat{\sigma}^2$ is adjusted as in equation (4) for the normal0 models. pVal cutoff indicates p-value cutoff used for model selection, where 1 respresented no model selection. Blue bars and points indicates model fitting, and pink indicates testing accuracy.

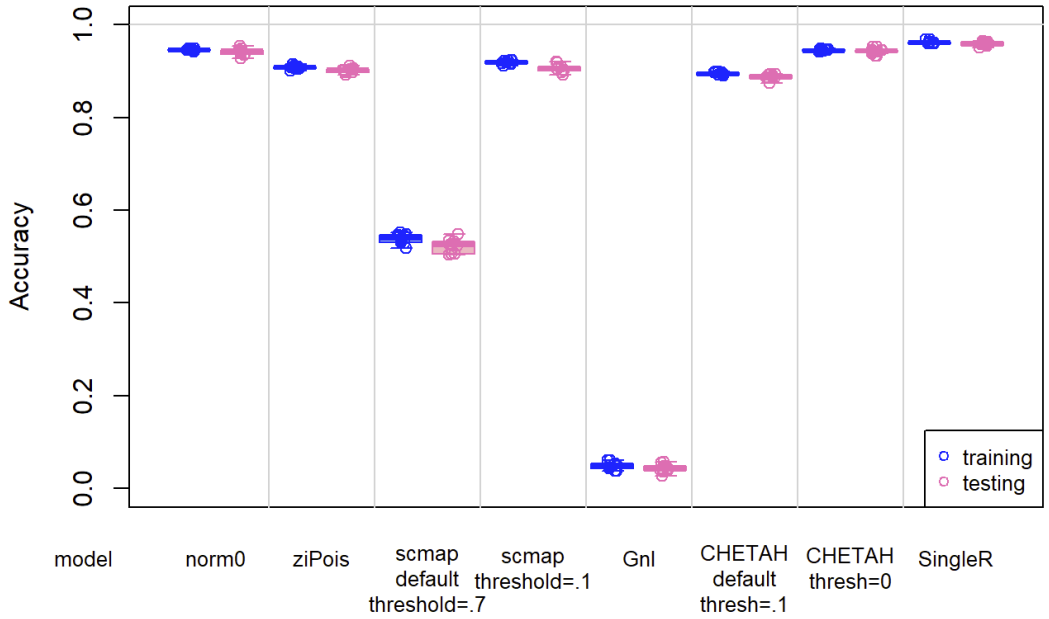

Figure 7: Comparing accuracy for model fitting and prediction accuracy between likelihood models and available methods with SMARTseq data. Points added on top of bar plots. The norm0 model shown here uses raw counts input and model selection with p-value cutoff $1e-4$. CHETAH and scmap are fitted twice using either their default or lowered threshold parameter. Blue bars and points indicates model fitting, and pink indicates testing accuracy.

The zi-Poisson model and the best performing norm0 model, which uses log2(cpm+1) input, model selection with p-value cutoff = $1e-4$, and adjusted $\hat{p_0}$ and $\hat{\sigma}^2$, were compared with other methods. The zi-Poisson model outperforms the default scmap and has similar performance as default CHETAH. The norm0 model outperforms both default scmap and default CHETAH. scmap and CHETAH with lowered threshold parameters, SingleR, and norm0 have prediction accuracy near 100% (*Figure 9*).

3.3.2 From SMARTseq to 10X

When training norm0 or zi-Poisson models with SMARTseq data and predict cell types of cells from 10X data, all versions of likelihood models have prediction accuracy near 100%. There is little room for improvement without including model selection and parameter adjustments (*Figure 10*).

The zi-Poisson model and the norm0 model, which uses $log2(cpm+1)$ input, model selection and no parameter adjustments, were compared with other methods. The likelihood models outperform

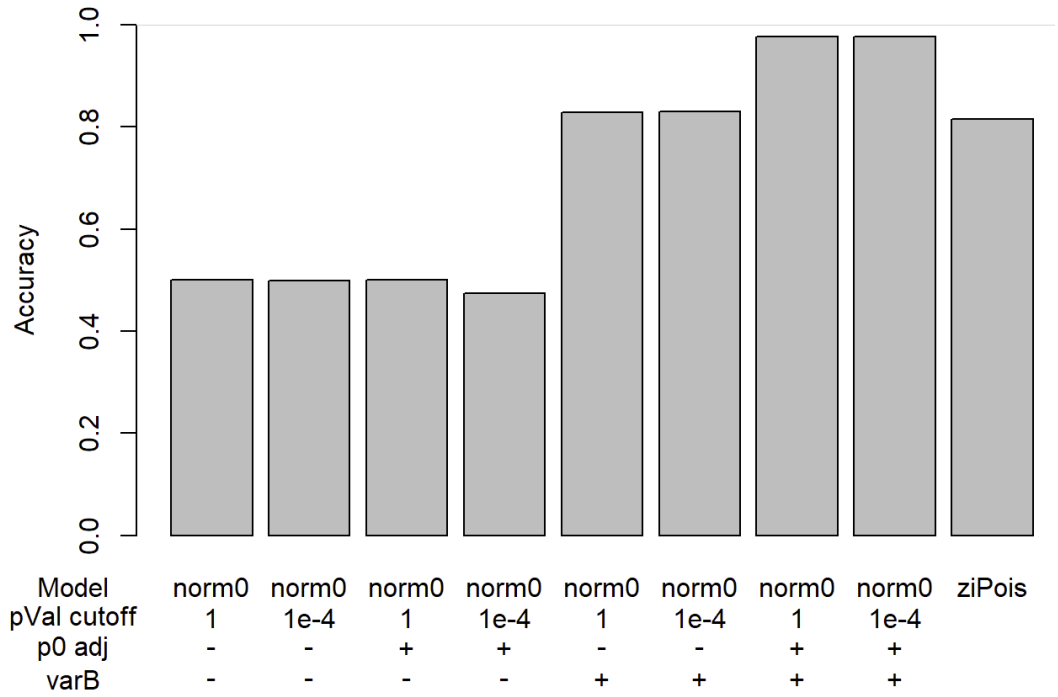

Figure 8: Comparison of cross platform prediction accuracy for likelihood models, using 10X as training data to predict SMARTseq cells. All norm0 models use $log_2(cpm+1)$ input. pVal cutoff indicates p-value cutoff used for model selection, where 1 represents no model selection. p0 adj indicates whether $\hat{p_0}$ is adjusted before prediction, basing on the mapping shown in Figure 3. VarB indicates whether $\hat{\sigma}^2$ is adjusted as in equation (4) for the normal0 models.

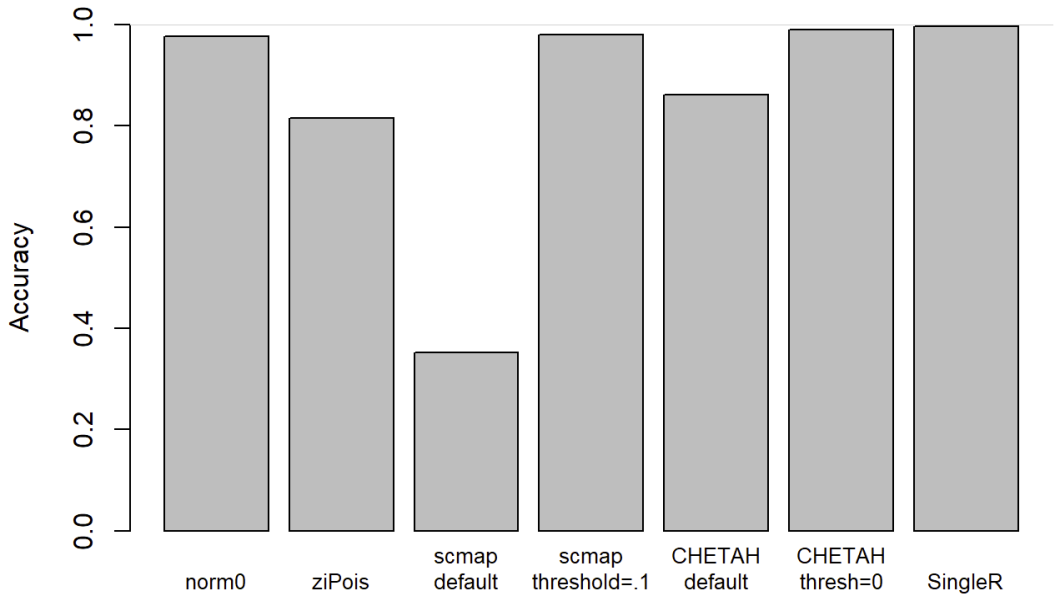

Figure 9: Comparison of cross platform prediction accuracy for available models, using 10X as training data to predict SMARTseq cells. The norm0 model uses $log_2(cpm+1)$ input, model selection with p-value cutoff = $1e-4$, and adjusted $\hat{p_0}$ and $\hat{\sigma}^2$. CHETAH and scmap are fitted twice using either their default (.7 for scmap and .1 for CHETAH) or lowered threshold parameter.

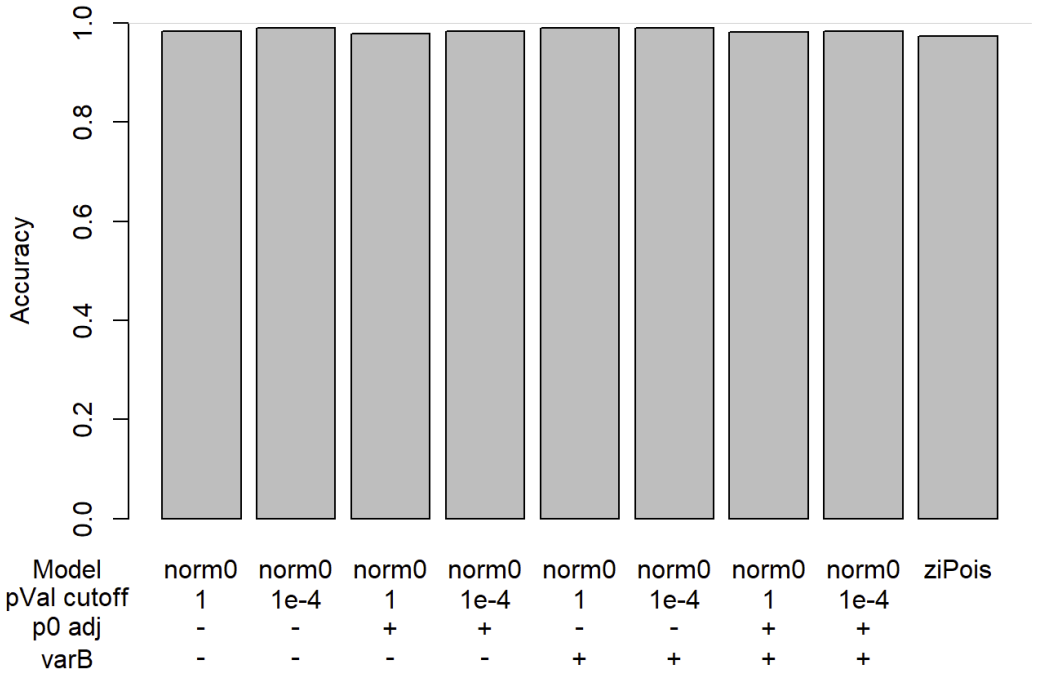

Figure 10: Comparison of cross platform prediction accuracy for likelihood models, using SMARTseq as training data to predict 10X cells. All norm0 models use $log_2(cpm+1)$ input. pVal cutoff indicates p-value cutoff used for model selection, where 1 represents no model selection. p0 adj indicates whether $\hat{p_0}$ is adjusted before prediction, basing on the mapping shown in *Figure 3*. VarB indicates whether $\hat{\sigma}^2$ is adjusted as in equation (4) for the normal0 models.

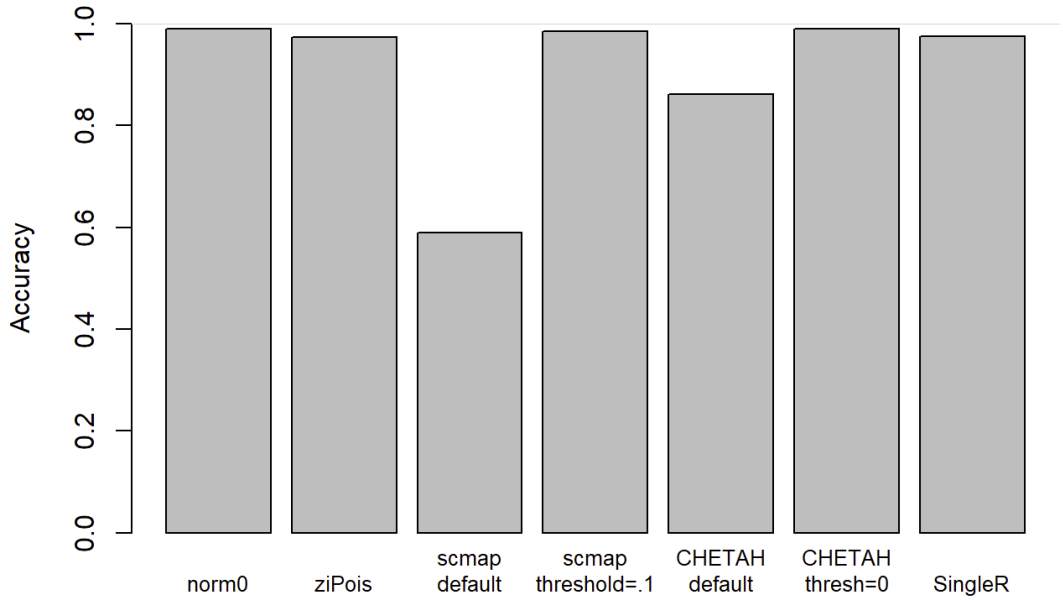

Figure 11: Comparison of cross platform prediction accuracy for available models, using SMARTseq as training data to predict 10X cells. The norm0 model uses $log_2(cpm+1)$ input, model selection with p-value cutoff = 1e-4, and no addiitonal adjustment for estimated parameters. CHETAH and scmap are fitted twice using either their default (.7 for scmap and .1 for CHETAH) or lowered threshold parameter.

both default scmap and default CHETAH. scmap and CHETAH with lowered threshold parameters, SingleR, norm0, and zi-Poisson all have prediction accuracy near 100% (*Figure 11*).

4 Discussion

We successfully built likelihood-based cell classification models that could transfer cell labels between datasets generated by different sequencing platforms. It was shown, that for both within dataset and cross platform prediction, our methods are among the best compared to the other available classification tools. To further verify these results, more datasets should be tested. Specifically, in addition to withholding test sets from the same dataset and additional cross platform datasets, external test sets generated with the same sequencing platform, but in different studies, should also be examined.

As model selection on the distributions of non-zero expressed gene in the norm0 models has shown to reduce over-fitting, similar model selection strategies can be implemented for the $\hat{p_0}$ parameters in norm0 models and both $\hat{\pi}$ and $\hat{\lambda}$ parameters in zi-Poisson models. This will further collapse the number of parameters estimated for each model. Additionally, this allows different cell types to potentially share estimated parameters, enabling the model to retain information when a gene is differentially expressed in some cell types, while reducing the noise when the gene was not differentially expressed in some other cell types. Thus the genes to be included in the model are not required to be markers for a specific cell type. For a dataset with K cell types, $5+2K$ parameters are used in GnI to model each gene; in contrast, our methods use up to $2K$ or $3K$ parameters, depending on specific model set-up for each gene, which reduces the risk of over-fitting.

One limitation of the current implementation of model selection is that the F tests are performed in an order that tested whether two cell types that have the closed estimated μ should be collapsed. If at one step, the null hypothesis is rejected, no further exploration is done for the cell types that are not compared. It is likely, that some of them should be collapsed to obtain a further reduced model. Alternative approaches for model selection should be explored.

Conceptually, having some cell types sharing estimated parameters will allow the model to include more genes without having to filter features to find a list of marker genes for each cell type. However, for cross datasets or platforms prediction, the feature selection step may still be necessary, since not all genes that carry information in one dataset still have useful information captured in another dataset.

Overall the correlation-based methods (scmap, CHETAH, and SingleR) are able to have reasonable or good results when predicting cross sequencing platforms. This implies that the selected genes by each method carry sufficient information for classification in both datasets. However, for scmap and CHETAH, the similarity score thresholds may need to be lowered to force the model to assign a cell type for a new cell, which is the limitation of correlation based methods. For SingleR, the model does not further summarize the training data into centroids for each cell type and uses only the top 80% of correlations calculated per cell type to propose cell labels, which makes it less sensitive to outliers in the training sets.

For cross platform prediction with the likelihood models, the norm0 models require both $\hat{p_0}$ and $\hat{\sigma}^2$ adjusted to obtain the best prediction results when training with 10X data and predicting cells from SMARTseq. Since more unobserved genes are expected in the 10X dataset due to the low coverage when sequencing a large number of cells, it is reasonable to reduce $\hat{p_0}$ before predicting cell labels for cells from the SMARTseq dataset, which has a higher coverage. On the other hand, the $\hat{\sigma}^2$ was adjusted according to equation (4), however $log_2(cpm+1)$ input was used to fit the model, thus the Poisson mean equals to variance assumption would not be applicable here. A more relevant way to adjust the $\hat{\sigma}^2$ is needed. Still, the need to inflate the $\hat{\sigma}^2$ implies that there is more variation in the non-zero expressions in the SMARTseq dataset than the observed variation in the 10X dataset.

Currently, our likelihood methods only give the name of the predicted cell type. With the calculated likelihood for each cell type, an additional score can be calculated to indicate how much confidence we have for this prediction. Moreover, our models are restricted to predict labels that are given in the training data. This can be extended in two ways: providing more than one possible cell type and couple with scores indicating confidence, and calling a cell to be a novel type that is not included in the training cells. Such modifications will make the models more applicable when possible novel cell types are expected in the new dataset.

5 References

1. Nguyen, A., Khoo, W. H., Moran, I., Croucher, P. I. & Phan, T. G. Single cell RNA sequencing of rare immune cell populations. *Frontiers in immunology* **9,** 1553 (2018).

2. Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nature communications* **10,** 1–11 (2019).

3. Tam, P. P. & Ho, J. W. Cellular diversity and lineage trajectory: insights from mouse single cell transcriptomes. *Development* **147** (2020).

4. Grabski, I. N. & Irizarry, R. A. A probabilistic gene expression barcode for annotation of cell-types from single cell RNA-seq data. *bioRxiv* (2020).

5. Zhang, A. W. *et al.* Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nature methods* **16,** 1007–1015 (2019).

6. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nature methods* **16,** 983–986 (2019).

7. Lieberman, Y., Rokach, L. & Shay, T. CaSTLe–classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PloS one* **13,** e0205499 (2018).

8. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods* **15,** 1053–1058 (2018).

9. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nature methods* **15,** 359–362 (2018).

10. De Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T. & Holstege, F. C. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic acids research* **47,** e95–e95 (2019).

11. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology* **20,** 163–172 (2019).

12. Bakken, T. E. *et al.* Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse. *bioRxiv* (2020).

13. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57,** 289–300 (1995).

14. McKenzie, A. T. *et al.* Brain cell type specific gene expression and co-expression network architectures. *Scientific reports* **8,** 1–19 (2018).