MAGE-SELECT: scRNA-seq Marker Gene Selection as a Multi-Level Bayesian Logistic Regression Problem

Mackenzie Kong-Sivert, Sanna Madan, and Elliott Sloate

December 2020

1 Introduction

Single cell RNA-sequencing (scRNA-seq) is a powerful technique which allows for profiling the transcriptomes of individual cells at an unprecedented scale and resolution. This technology allows us to peek into and better understand the biology underlying states such as disease and development. A key challenge in analyzing scRNA-seq is identifying "marker" genes that can uniquely identify cell (sub)types. Cell (sub)types have hierarchical relationships that need to be accounted for, adding to the technical complexity of this challenge. In this paper, we present a novel, data-driven approach for identifying sets of marker genes that are predictive of cell (sub)types, based on a multi-level Bayesian logistic regression. We term this approach MAGE-SELECT (MArker GEne SELECTion).

2 Previous Work

Numerous approaches [1, 2, 6, 5] have been developed in the recent years to solve this key challenge, incorporating statistics, machine learning, and combinatorial optimization techniques. Some notable approaches will be discussed here. One algorithm, COMET, selects markers to identify cell populations via a non-parametric statistical framework. scGene-Fit does the same using label-aware compressive classification methods. OnClass, on the other hand, uses the Cell Ontology graph to infer relationships between cell (sub)types, regardless of whether a given cell type was present in the training data. Finally, a recent approach, Venice, uses a new metric to identify genes which are up- or down-regulated, or in transitional states, to identify marker genes.

3 Data

We analyzed data from two different scRNA-seq sources. Both data sources are matrices where rows represent cells, columns represent genes, and the values represent gene expression levels for a particular cell. We first worked on CITE-seq data from Stoeckius *et al* [3]. CITE-seq is a high-throughput method which profiles both the transcriptome and cell surface proteins (epitopes) in a single-cell readout. This data was 8,617 cells x 500 genes, taken from human and mouse cord blood mononuclear cells. The second source of data came from Zeisel et al, consisting of 3,005 cells x 4,000 genes taken from mouse cortical cells [8].

4 Methods

4.1 Approach

4.1.1 Preprocessing and Clustering

First, we carried out preprocessing of the scRNA-seq matrices. The data were normalized using an L1-norm. Next, we assigned each cell to a "meta-cluster". These clusters, while not reflecting the final cell subtypes of the cells, instead reflect cell subtypes which share similar features. To initially cluster the CITE-seq and Zeisel data, we performed a k-means clustering. Since CITE-seq consisted of 13 unique cell subtypes and Zeisel consisted of 7, we chose to compute 8 clusters for the CITEseq data and 4 for the Zeisel data.

Computing these clusters was an essential step, as treating different cell subtypes as wholly distinct would throw away useful information and potentially result in highly variable parameter estimates for scarce cell subtypes [7]. We instead recognize that there is a hierarchical relationship between cell subtypes, where some individual cell subtypes can be feasibly grouped together.

4.1.2 Regression Model

In order to use this information, we chose to perform a multi-level logistic regression, where each cell's meta-cluster was used as a random intercept effect. In regression problems, random effects are used when individual units are nested together [4]. We explicitly grouped cells together by meta-cluster, and expected the meta-cluster assignments to have a large effect on the variance of our outcome, (*i.e.*, cells that have the same meta-cluster are more likely to be from the same cell type). Therefore, a random effect is an ideal way to model this relationship. This random effect allows us to differentiate meta-clusters, while acknowledging that there is also a potential relationship *between the meta-clusters*. Thus, the intercept estimates for meta-clusters with low prevalence in the data will be pulled towards the mean intercept value [7].

In addition to using a multi-level model, we hypothesized that a Bayesian model, where model parameters are given prior distributions, would be well-

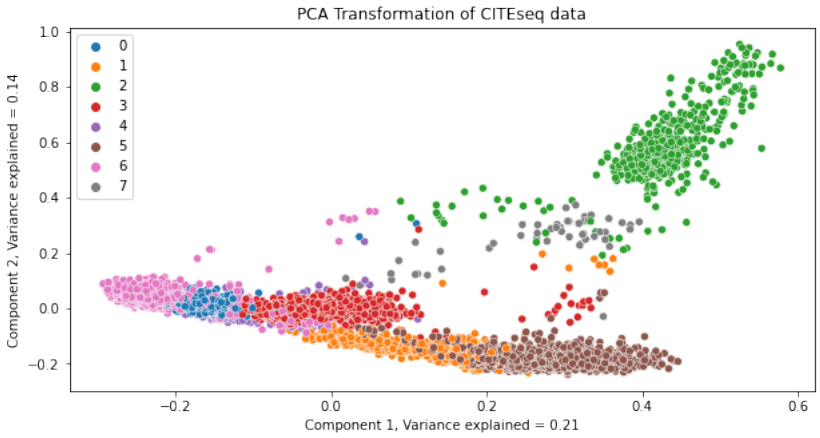

Figure 1: PCA transformation of CITE-seq data

suited for this problem. For marker gene selection, we believe that most genes should have no effect on cell type, and instead only a few should be considered significant. A simple way to model this is by giving each β coefficient (corresponding to each gene) a prior centered tightly around 0. With this prior, only a small number of genes will be selected as significant (that is, we can confidently say the β value is greater than 0). In our case, we chose a Laplace prior with $\mu = 0$ and $b = 1$. However, this is just one of many priors one could choose. If one has better prior knowledge of cell-gene relationships, different priors could be chosen for different β values.

Our approach can be summarized as follows. First, we selected h meta-clusters for which each cell is assigned to. Then, we selected k genes of interest, which we chose as the genes that have the highest expression variance between cells. Then, for each cell type m, where $1 \leq m < M$, we performed a multi-level Bayesian logistic regression, where the intercept α was modeled as a random effect (in our case, coming from a Normal distribution), and each β coefficient had a prior distribution (in our case, all βs are assigned a Laplace prior). The formula for each regression is thus:

$$\begin{aligned} P(cell\ subtype = m) &\sim Bernoulli(p_i) \\ logit(p_i) &= \alpha_{h,i} + \beta_1 x_i + \beta_2 x_i + ... + \beta_k x_i \\ \beta_{1...k} &\sim Laplace(0, 1) \\ \alpha_{h,\cdot} &\sim Normal(\mu_\alpha, \sigma_\alpha^2) \\ \mu_\alpha &\sim Normal(0, 2) \\ \sigma_\alpha &\sim HalfNormal(1) \end{aligned} \tag{1}$$

Note that in a non-Bayesian multi-level regression, $\alpha_{h,\cdot}$ would still have a Normal prior because it is a random effect, but with a fixed mean, it would perhaps be 0. Because this is a Bayesian regression, the parameters of our

priors have priors! After each regression, we took the β parameters with a 95% Highest Posterior Density Interval entirely above 0 as marker genes for that cell type. Since we performed separate regressions, this is highly parallelizable. However, the Python package PyMC3 uses all the cores on a standard PC, so we could not parallelize it for our task. In theory however, one could run a regression for all the cell subtypes at once.

5 Results

5.1 Algorithm Output

For a given number of candidate marker genes, our algorithm will potentially assign each cluster marker genes. Table 1 shows an example output for the Zeisel data (7 cell subtypes) with 10 candidate marker genes.

Cell Type (Name)	Marker Genes
1 (Astrocytes ependymal)	7,8
2 (Endothelial-mural)	5,10
3 (Interneurons)	8
4 (Microglia)	8,9
5 (Oligodendrocytes)	5
6 (Pyramidal CA1)	-
7 (Pyramidal SS)	6,7

Table 1: Marker Genes Assigned To Each Cell Type

Here we see that out of the 10 candidate marker genes, at most two are assigned to any given cell type, and each cell type (besides type 6) can be assigned its own marker gene (i.e. 1: 7, 2: 10, 3: 8, 4: 9, 5: 5, 7: 6). While cell type 6 was not assigned a marker gene in this case, we found that all types were assigned marker genes for M = 20, 25, 30 using the Zeisel data. For CITE-seq data, all cell subtypes (13 total) were assigned marker genes for M = 25, 30. In Figure 2 is a heatmap displaying the expression levels of our selected marker genes (for k = 10) across the different cell subtypes in the Zeisel data. We see that gene 5 is significantly more expressed in cluster 5 than in other clusters, which is the expected indicator of a marker gene.

5.2 Runtime Performance

We first analyzed the runtime based on the number of cells we included in our dataset. We observed that for both the number of cells in the data, as well as the number of candidate marker genes, we see a linear increase in the run time of a single regression as either of these parameters increase.

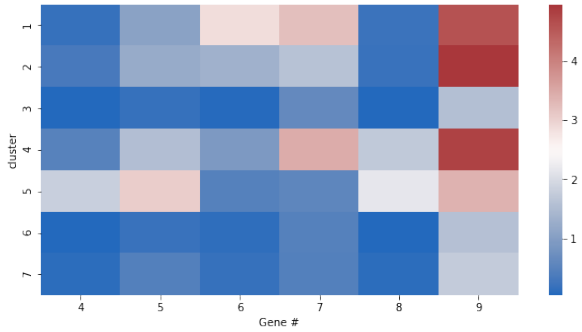

Figure 2: Expression of marker genes across different cell subtypes

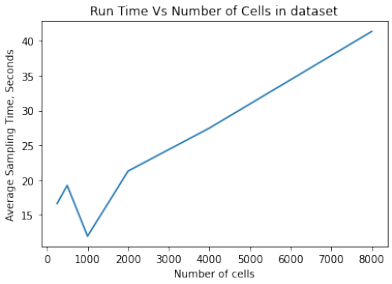

Figure 3: Run time vs. the number of cells in the dataset, number of marker gene candidates = 15

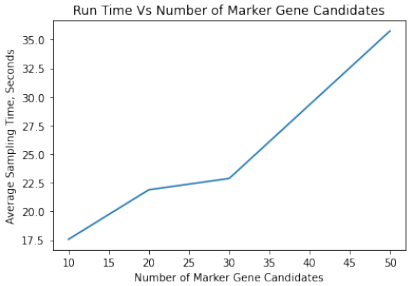

Figure 4: Run time vs. the number of marker genes candidates, number of cells = 500

5.3 Accuracy

To validate our marker genes, we followed the approach employed by Dumitrascu *et al.*, by using a nearest centroid classifier to predict class labels using only the marker genes identified by our approach [2]. Table 2 shows how our algorithm performs compared to scGene-Fit. Because for a desired number of marker genes candidates k, our algorithm will not necessarily assign all candidate marker genes to cells (unlike scGene-Fit), we include the actual number of marker genes our algorithm actually assigned in parentheses.

Our method outperforms scGeneFit on the CITEseq data, with a significant performance boost when looking at 30 candidate marker genes. scGeneFit outperforms MAGE-SELECT on the Zeisel data, where neither algorithm has a strong performance.

Model	Data	Marker Genes (total assigned)	Accuracy
—	CITEseq	500	95.8%
scGeneFit	CITEseq	20	79.6%
MAGE-SELECT	CITEseq	20 (13)	**83.2**%
scGeneFit	CITEseq	25	89.0%
MAGE-SELECT	CITEseq	25 (17)	**91.1**%
scGeneFit	CITEseq	30	83.2%
MAGE-SELECT	CITEseq	30 (22)	**92.3**%
—	Zeisel	4000	36.9%
scGeneFit	Zeisel	20	37.3%
MAGE-SELECT	Zeisel	20 (14)	**37.6**%
scGeneFit	Zeisel	25	**38.3**%
MAGE-SELECT	Zeisel	25 (17)	37.7%
scGeneFit	Zeisel	30	**38.3**%
MAGE-SELECT	Zeisel	30 (22)	37.6%

Table 2: Accuracy scores for scGeneFit and MAGE-SELECT for varying numbers of marker genes

6 Discussion

We believe that the usage of Bayesian priors provides a unique method for determining the most important genes for a given cell type, and that multi-level models are a useful tool for capturing the hierarchical relationships between cell subtypes. These methods can be used together with any clustering algorithm for robust scRNA-seq marker gene detection. Based on our model's performance in recovering cell subtypes from marker genes, we conclude that MAGE-SELECT improves upon some of the already established techniques for scRNA-seq marker gene detection.

6.1 Future Work

There are several exciting future directions this work could take. First, we can improve the meta-cluster generation step, via integrating *a priori* biological knowledge and/or employing a different clustering algorithm than k-means. In addition, it would be beneficial to test MAGE-SELECT on additional scRNA-seq/CITE-seq datasets and compare its performance against other existing methods.

Finally, as discussed earlier, our algorithm runs rather slowly and is not parallelized. It is in theory parallelizable, but the Python package PyMC3 which we integrated used all cores. The regression can and ideally should be parallelized for all cell subtypes to speed up the analysis.

7 Code availability

The code for MAGE-SELECT can be found at https://github.com/smadan20/single-cell-marker-gene-detection. The CITE-seq data were from [3] and the Ziesel data were from [8].

References

[1] Conor Delaney, Alexandra Schnell, Louis V Cammarata, Aaron Yao-Smith, Aviv Regev, Vijay K Kuchroo, and Meromit Singer. Combinatorial prediction of marker panels from single-cell transcriptomic data. *Molecular systems biology*, 15(10):e9005, 2019.

[2] Bianca Dumitrascu, Soledad Villar, Dustin G. Mixon, and Barbara E. Engelhardt. Optimal marker gene selection for cell type discrimination in single cell analyses. *bioRxiv*, 2019.

[3] William Stephenson Brian Houck-Loomis Pratip K Chattopadhyay Harold Swerdlow Rahul Satija Marlon Stoeckius, Christoph Hafemeister and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865, 2017.

[4] Micah Mumper. Multilevel modelling, Jan 2017.

[5] Hy Vuong, Thao Truong, Tan Phan, and Son Pham. Venice: A new algorithm for finding marker genes in single-cell transcriptomic data. *bioRxiv*, 2020.

[6] Sheng Wang, Angela Oliveira Pisco, Jim Karkanias, and Russ B. Altman. Unifying single-cell annotations based on the cell ontology. *bioRxiv*, 2019.

[7] Thomas Wiecki and Danne Elbers. Glm: Hierarchical linear regression¶.

[8] Amit Zeisel, Ana B. Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler, and Sten Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.