EDClust: An EM-MM hybrid method for cell clustering in population-level single cell RNA sequencing

Xin Wei, MSPH
Department of Biostatistics and Bioinformatics, Emory University

INTRODUCTION

- Single-cell RNA sequencing (scRNA-seq) technology has revolutionized genomics research.
- With the increasing application of scRNA-seq in larger scale studies, people face the problem of cell clustering when the scRNA-seq data are from more than one subject.
- One challenge in analyzing such data is the subject-specific systematic variations, which may have significant impacts on the clustering accuracy.
- Existing methods addressing such effect suffered from several limitations.
- We develop a novel statistical method named "**EDClust**" for scRNA-seq cell clustering when data are from multiple subjects.

METHODS

Data model:

- Sequence counts Y follows a Dirichlet-Multinomial mixture distribution:
  1. A cell type label $W_{li} \in \{1,2,...,K\}$ is assigned to cell i in subject l with probability $P(W_{li} = k) = \pi_{lk}$
  2. Given the cell label (i.e., $W_{li} = k$), Y_{li} will be generated from a Multinomial distribution by $Y_{li} \sim \text{Multinomial}(T_{li}, p_{li})$.
  3. p_{li} follows a cell-type specific prior distribution $\text{Dirichlet}(\alpha_{lk}) = \text{Dirichlet}(\alpha_{lk1}, \alpha_{lk2},..., \alpha_{lkJ})$.
- The overall effect α can be expressed as the sum of cell type effect α_0 and subject effect δ : $\alpha = \alpha_0 + \delta > 0$.

Algorithm:

- Expectation–Maximization (EM) algorithm is derived to maximize the observed data likelihood and obtain posterior probabilities for cell type assignment W_{li}.

$$\text{E-step:} \quad \mu_{lik}^{(t)} = \frac{\pi_{lk}^{(t)} P\left(Y_{li}|T_{li}, \alpha_{0k}^{(t)} + \delta_{lk}^{(t)}\right)}{\sum_{k'}^{K} \pi_{lk'}^{(t)} P\left(Y_{li}|T_{li}, \alpha_{0k'}^{(t)} + \delta_{lk'}^{(t)}\right)}$$

$$\text{M-step:} \quad \pi_{lk}^{(t+1)} = \frac{\sum_{i=1}^{I_l} \mu_{lik}^{(t)}}{I_l}$$

- Within the M-step of the EM, Minorize-Maximization (MM) algorithm is derived to update the cell type effect α_0 and subject effect δ.

$$\delta_{lkj}^{(t,n+1)} = \left(\sum_{c_{2lj}} \frac{s_{lkjc}^{(t)} \delta_{lkj}^{(t,n)}}{\alpha_{0kj}^{(t,n)} + \delta_{lkj}^{(t,n)} + c_{2lj}}\right) \Bigg/ \left(\sum_{c_{1l}} \frac{r_{lkc}^{(t)}}{\left\|\alpha_{0k}^{(t,n)} + \delta_{lk}^{(t,n)}\right\|_1 + c_{1l}}\right)$$

$$\alpha_{0kj}^{(t,n+1)} = \left(\sum_{l=1}^{L} \sum_{c_{2lj}} \frac{s_{lkjc}^{(t)} \alpha_{0kj}^{(t,n)}}{\alpha_{0kj}^{(t,n)} + \delta_{lkj}^{(t,n)} + c_{2lj}}\right) \Bigg/ \left(\sum_{l=1}^{L} \sum_{c_{1l}} \frac{r_{lkc}^{(t)}}{\left\|\alpha_{0k}^{(t,n)} + \delta_{lk}^{(t,n)}\right\|_1 + c_{1l}}\right)$$

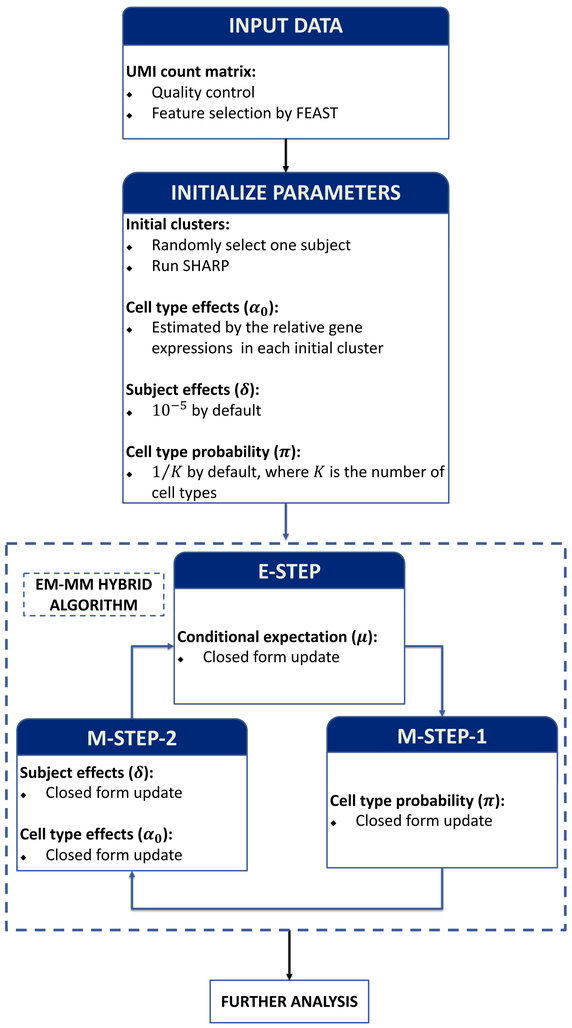

Figure 1. Summary of the EDClust algorithm.

RESULTS

- We design a series of simulation studies to evaluate the performance of EDClust and compare it to several competing methods. We evaluate the methods when data have different levels of subject specific effects (low, medium, and high), and with different sample size selections (5, 10, 15). EDClust constantly achieves the highest average adjusted Rand index (ARI).
- We benchmark EDClust and other methods on four real scRNA-seq datasets. For three out of four datasets, EDClust has the best performance, and the performance improvement can be significant. For example in the Mouse Retina data, EDClust has the mean ARI of 0.87, while the second best performer (Harmony+SC3) only has the ARI of 0.70. In the Mouse Lung data, EDClust performs slightly worse than BAMM-SC and Harmony+SC3.

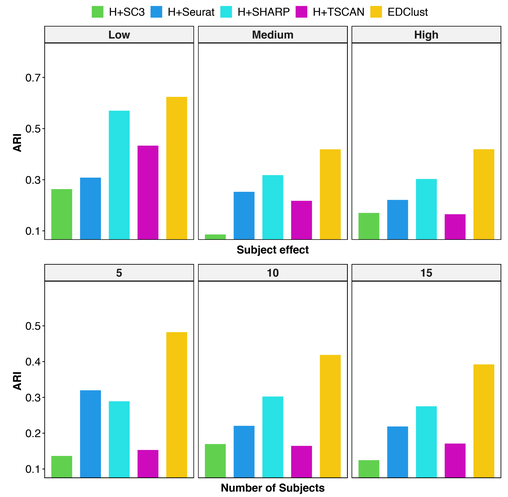

Figure 2. Bar plots of average ARIs for EDClust and competing clustering methods across over 20 simulations, where "H +" indicates that the simulation data are processed by Harmony to remove the subject effects.

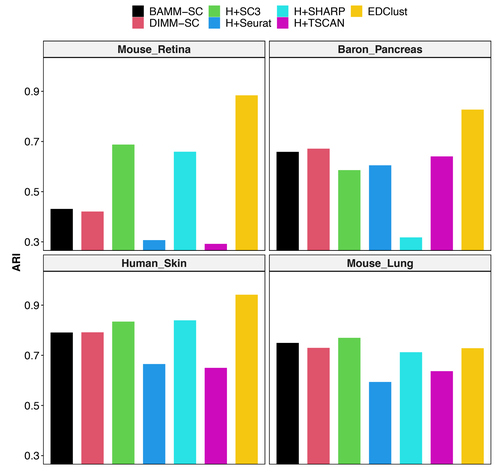

Figure 3. Bar plots of average ARIs for EDClust and competing clustering methods across over 50 clustering results. "H +" represents that the clustering methods are implemented on the real datasets which have been processed by Harmony to remove the subject effects.

CONCLUSIONS

We develop EDClust for cell clustering in multi-subject scRNA-seq data. We model the sequence read counts by a mixture of Dirichlet-Multinomial distributions and design an EM-MM hybrid algorithm for model-based clustering. EDClust mainly has the following advantages:

1. EDClust describes data heterogeneity among multiple subjects.
2. Utilizing the shared information among subjects, EDClust clusters all the cells from all subjects simultaneously, which improves the accuracy of cell clustering.
3. Most of the clustering methods require several pre-processing steps, while EDClust offers a one-stop service that can be directly applied to raw count data.
4. EDClust quantifies clustering uncertainty with the probability that each cell belongs to a given cluster, contributing to further statistical inference and biological interpretation.

EMORY ROLLINS SCHOOL OF PUBLIC HEALTH

Acknowledgement

Acknowledgement to Hao Wu, PhD, for his expertise and continuous support.
Acknowledgement to Zhaohui (Steve) Qin, PhD, for his invaluable comments.
Acknowledgement to Ziyi Li, PhD, for her assistance and insightful suggestions.

References

1. Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell, 161(5):1202–1214, 2015.
2. Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. Cell systems, 3(4):346–360, 2016.
3. Zhe Sun, Li Chen, Hongyi Xin, Yale Jiang, Qianhui Huang, Anthony R Cillo, Tracy Tabib, Jay K Kolls, Tullia C Bruno, Robert Lafyatis, et al. A bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies. Nature communications, 10(1):1–10, 2019.
4. Shibiao Wan, Junil Kim, and Kyoung Jae Won. Sharp: hyperfast and accurate processing of single-cell rna-seq data via ensemble random projection. Genome research, 30(2):205–213, 2020.
5. Kenong Su, Tianwei Yu, and Hao Wu. Accurate feature selection improves single-cell rna-seq cell clustering. Briefings in Bioinformatics, 2021.