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:

1. Sequence counts \( Y \) follows a Dirichlet-Multinomial mixture distribution:
   - A cell type label \( W_i \in \{1,2,...,K\} \) is assigned to cell \( i \) in subject \( j \) with probability \( P(W_{ij}=k) \).
   - The cell label \( Y_{ij} \) will be generated from a Multinomial distribution with \( Y_{ij} \sim \text{Multinomial}(\Theta_j) \).
   - \( \Theta_j \) follows a cell-type specific prior distribution \( \text{Dirichlet}(\Theta_j) = \text{Dirichlet}(\alpha_{01},\alpha_{02},...\alpha_{0K}) \).
   - The overall effect \( \alpha \) can be expressed as the sum of cell type effect \( \alpha_k \) and subject effect \( \delta \): \( \alpha = \alpha_k + \delta \).

Algorithm:

• Expectation–Maximization (EM) algorithm is derived to maximize the observed data likelihood and obtain posterior probabilities for cell type assignment \( W_i \).
• E-step: \( \mu_{ij}(\alpha) = \frac{\sum_{k=1}^{K} \gamma_{ij}(k)}{\sum_{k=1}^{K} \gamma_{ij}(k) + \sum_{l=1}^{L} \gamma_{lj}(l)} \).
• M-step: \( \alpha \leftarrow \alpha \leftarrow \mu_{ij}(\alpha) \).
• Within the M-step of the EM, Minorize-Maximization (MM) algorithm is derived to update the cell type effect \( \alpha_k \) and subject effect \( \delta \): \( \mu_{ij}(\alpha) = \frac{\sum_{k=1}^{K} \gamma_{ij}(k)}{\sum_{k=1}^{K} \gamma_{ij}(k) + \sum_{l=1}^{L} \gamma_{lj}(l)} \).

INPUT DATA

• UMI count matrix:
  - Quality control
  - Feature selection by FEAST

INITIALIZE PARAMETERS

• Initial clusters:
  - Randomly select one subject
  - Run SHARPI
• Cell type effects (\( \alpha_k \)):
  - Estimated by the relative gene expression in each initial cluster
• Subject effects (\( \delta \)):
  - \( 1/\text{subject} \) by default
• Cell type probability (\( \alpha_k \)):
  - \( 1/\text{subject} \) by default, where \( K \) is the number of cell types

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.

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-step 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.

Acknowledgement

Acknowledgement to Hao Wu, PhD, for his expertise and continuous support. Acknowledgement to Ziyi Li, PhD, for her assistance and insightful suggestions.

References