scAnnotate

an automated cell type annotation tool for single-cell RNA-sequencing data

University of Victoria

Xiangling Ji, Danielle Tsao , Kailun Bai, Min Tsao, Li Xing, Xuekui Zhang

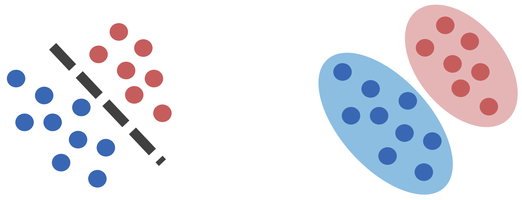
Department of Mathematics and Statistics, March 2023

Introduction

- ***Cell annotation***: the labelling of cell types in biological samples
- ***Single-cell RNA sequencing (scRNA-seq)***: a recently developed technology to examine gene information at the single-cell level
- Challenges of scRNA-seq data: sparsity and high-dimensionality

Q: How can we annotate cells based on their scRNA-seq data?

- Two approaches in machine learning:



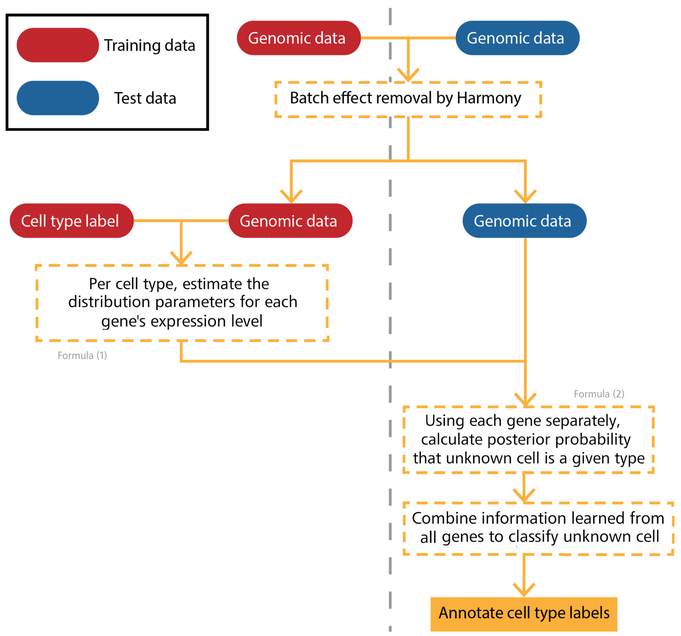
1. Discriminative models: draw decision boundaries

2. Generative models: model distribution of the data

- However, existing methods do not consider dropout (excessive amount of zeros), a key characteristic of scRNA-seq data!

A: We introduce **scAnnotate**, a cell annotation tool for scRNA-seq data with a generative approach.

Figure 1. Work flow of scAnnotate at a glance



Methods

- Given prior probability π_i that unknown cell C is type i, the prior distribution for gene j's expression level X_j (learned from training data) is $F_j = \sum \pi_i F_{ij}$ where

$$F_{ij} = p_{ij}F^0 + (1 - p_{ij})F_{ij}{}^+ \quad (1)$$

- p_{ij}: mixing proportion; distributions F^0 : degenerate at 0 and $F_{ij}{}^+$: support $(0, \infty)$.
- For observed test data value x_j of X_j, the posterior probability that cell C is type i is

$$q_{ij} = P(type\ i\ |X_j = x_j) = \frac{P(X_j = x_j \mid type\ i)\ P(type\ i)}{P(X_j = x_j)} \quad (2)$$

- Classify cell C such that combiner function s_i of q_{ij} is maximized over all types i.

Results

1. Cross-platform classification performance

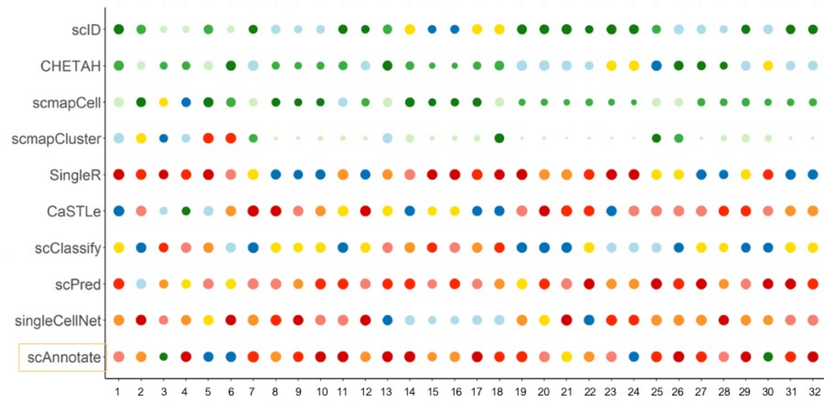


Figure 2. Performance of scAnnotate and existing cell annotation methods [7-14] when scRNA-seq training and test datasets (PBMC [1] [2] and lung cancer cells [3]) are generated on different platforms

scAnnotate is a **top 3** ranked method for **cross-platform cell classification.**

2. Complementary predictive strength to other methods

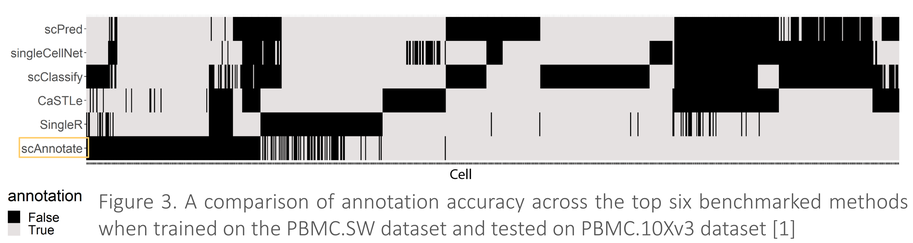


Figure 3. A comparison of annotation accuracy across the top six benchmarked methods when trained on the PBMC.SW dataset and tested on PBMC.10Xv3 dataset [1]

scAnnotate is often correct when the top 5 *discriminative* methods are not.

3. Cross-species classification performance

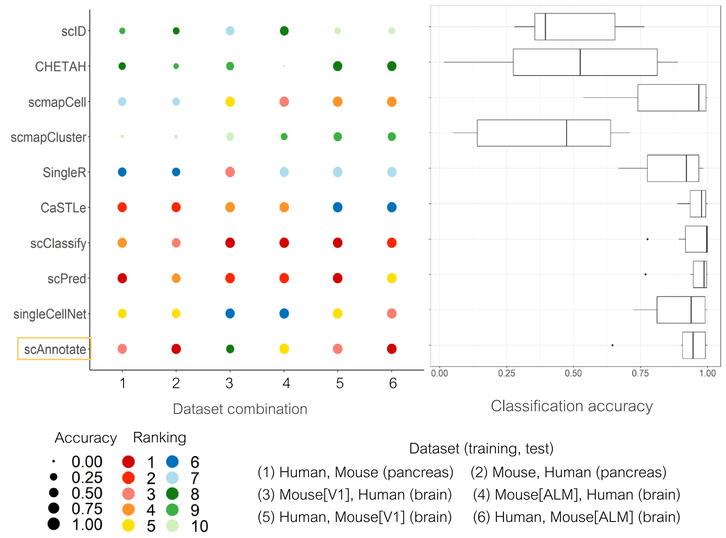


Figure 4. Performance of scAnnotate and existing cell annotation methods on six combinations of cross-species scRNA-seq datasets (human/mouse brain [4] [5] and pancreas [6] samples)

scAnnotate consistently ranks in the **top 5** best performing methods for **cross-species cell classification**.

Conclusions

- scAnnotate provides top-tier cell classification performance.
- scAnnotate is likely a key player in a mega ensemble model with competing discriminative methods.

References

1. Ding et al. BioRxiv (2019): 632216.
2. Tian et al. Nature methods 16.6 (2019): 479-487.
3. Abdelaal et al. Genome biology 20.1 (2019): 1-19.
4. Tasic et al. Nature 563.7729 (2018): 72-78.
5. Hodge et al. Nature 573.7772 (2019): 61-68.
6. Baron et al. Cell systems 3.4 (2016): 346-360.
7. Lieberman, Y., L. Rokach, and T. Shay. PloS one 13.10 (2018): e0205499.
8. Boufea, K., S. Seth, and N. Batada. Science 23.3 (2020): 100914.
9. Lin et al. Molecular systems biology 16.6 (2020): e9389
10. Tan, Y. and P. Cahan. Cell systems 9.2 (2019): 207-213.
11. Alquicira-Hernandez et al. Genome biology 20.1 (2019): 1-17.
12. Aran et al. Nature immunology 20.2 (2019): 163-172.
13. de Kanter et al. Nucleic acids research 47.16 (2019): e95-e95.
14. Kiselev, V., A. Yiu, and M. Hemberg. Nature methods 15.5 (2018): 359-362.

Acknowledgements

University of Victoria

This research was supported by the Jamie Cassels Undergraduate Research Awards, University of Victoria. Supervised by Dr. Zhang, Dr. Tsao and Dr. Xing.