Ziyi lab on Statistical Genomics at MDA

THE UNIVERSITY OF TEXAS
MD Anderson
~~Cancer~~ Center
Making Cancer History®

Ziyi Li

Department of Biostatistics
The University of Texas MD Anderson Cancer Center
zli16@mdanderson.org

Directions of our lab

METHODOLOGY DEVELOPMENT FOR

- Single cell technology

  Novel cell detection, longitudinal design, population-scale analysis
- Spatial omics data analysis

  Cell type mediation analysis, cell annotation
- TCR-seq data analysis

  Longitudinal design, sequence interpretation
- Problems in clinical data analysis

  Risk estimation and prediction

COLLABORATIVE RESEARCH

Single cell technology

- Most of the biological experiments are performed on “bulk” samples, which contain a large number of cells (millions).
- The “bulk” data measure the average signals (gene expression, TF binding, methylation, etc.) of many cells.
- The bulk measurement ignores the inter-cellular heterogeneities:
  - Different cell types.
  - Variation among the same cell type.

Single cell technology

- Single-cell biology: the study of individual cells.
- The cells are isolated from multi-cellular organism. Experiment is performed for each cell individually.
- Provides more detailed, higher resolution information. High-throughput experiments on single cell is possible.

- Different types of sequencing: DNA-seq, ATAC-seq, BS-seq, RNA-seq, multi-omics

Single Cell RNA Sequencing Workflow

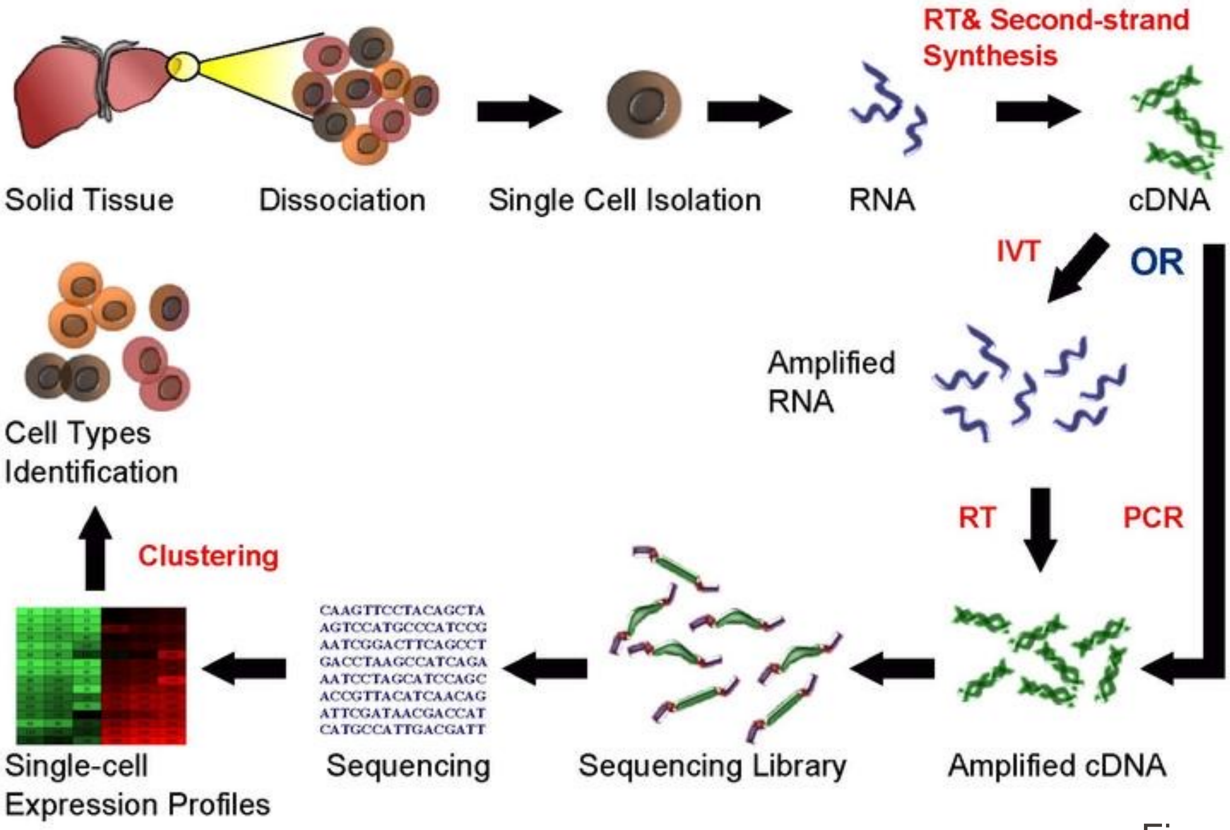


Figure source: wikipedia

Single cell RNA-seq (scRNA-seq)

- The **most active** in the single cell field.

- **Scientific goals:**
  - Composition of different cell types in complex tissues.
  - New/rare cell type discovery.
  - Gene expression, alternative splicing, allele-specific expression at the level of individual cells.
  - Transcriptional dynamics (pseudotime construction).
  - Above can be investigated and compared spatially, temporally, or under different biological conditions.

- **Technology:**
  - Plate-based methods (Smart-seq, Smart-seq2, CEL-seq)
  - Droplet-based methods (Drop-seq, inDrop, 10x genomics)

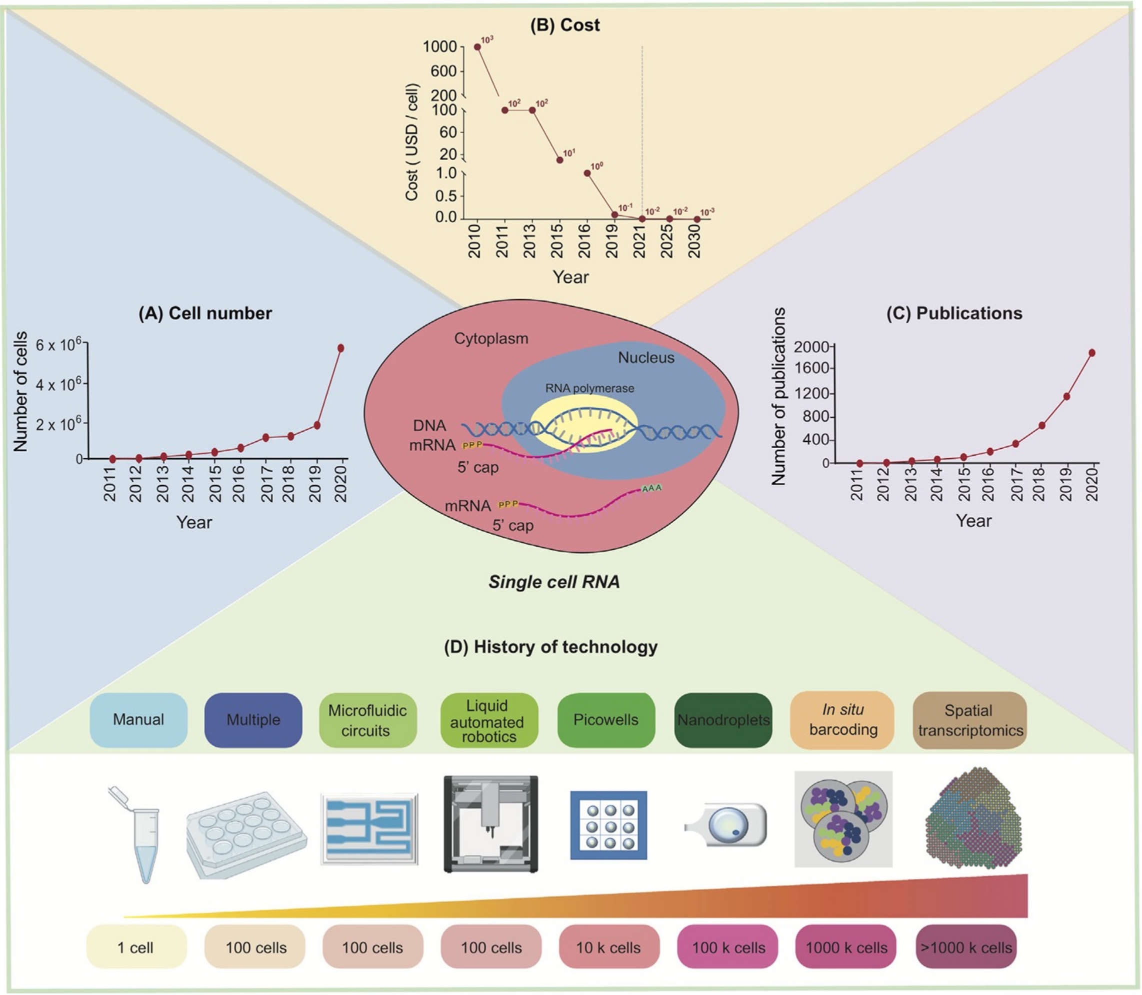
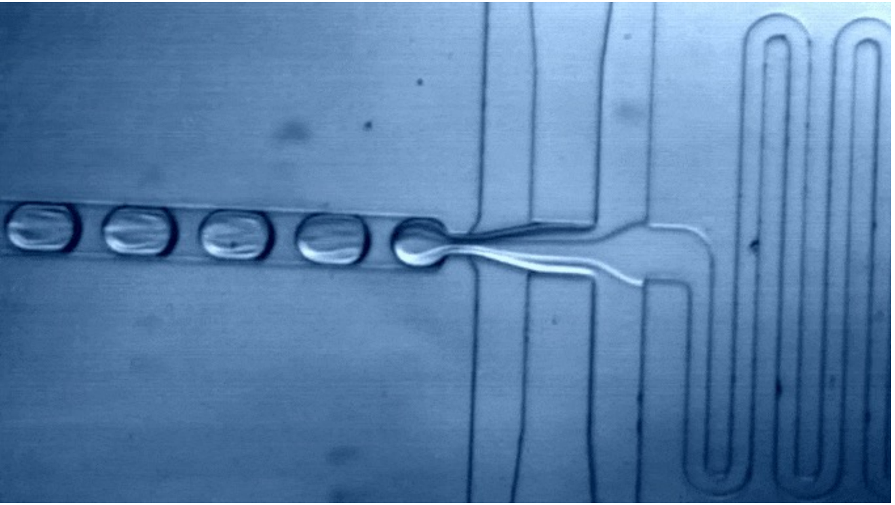


Figure source: Jovic, Dragomirka, et al. "Single-cell RNA sequencing technologies and applications: A brief overview." *Clinical and Translational Medicine* 12.3 (2022): e694.

Illustration of Drop-let based technology



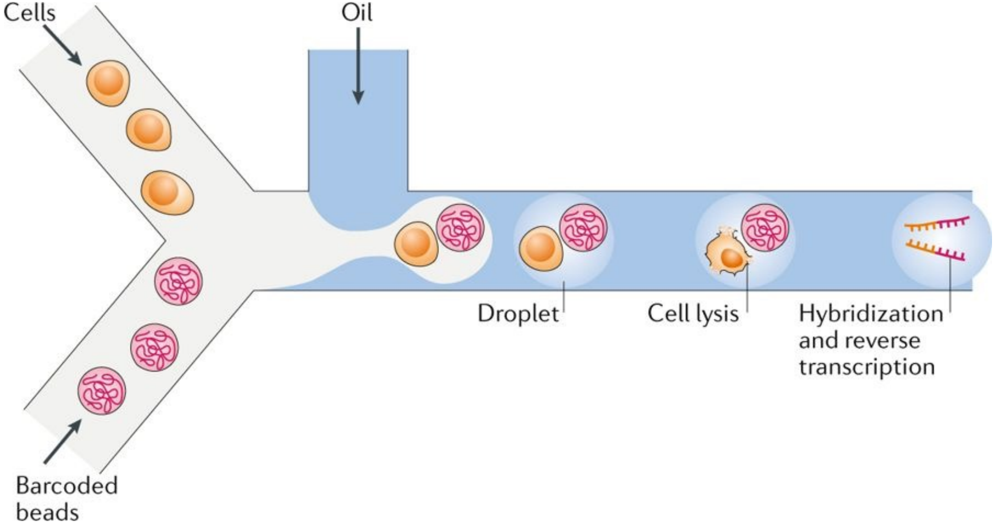
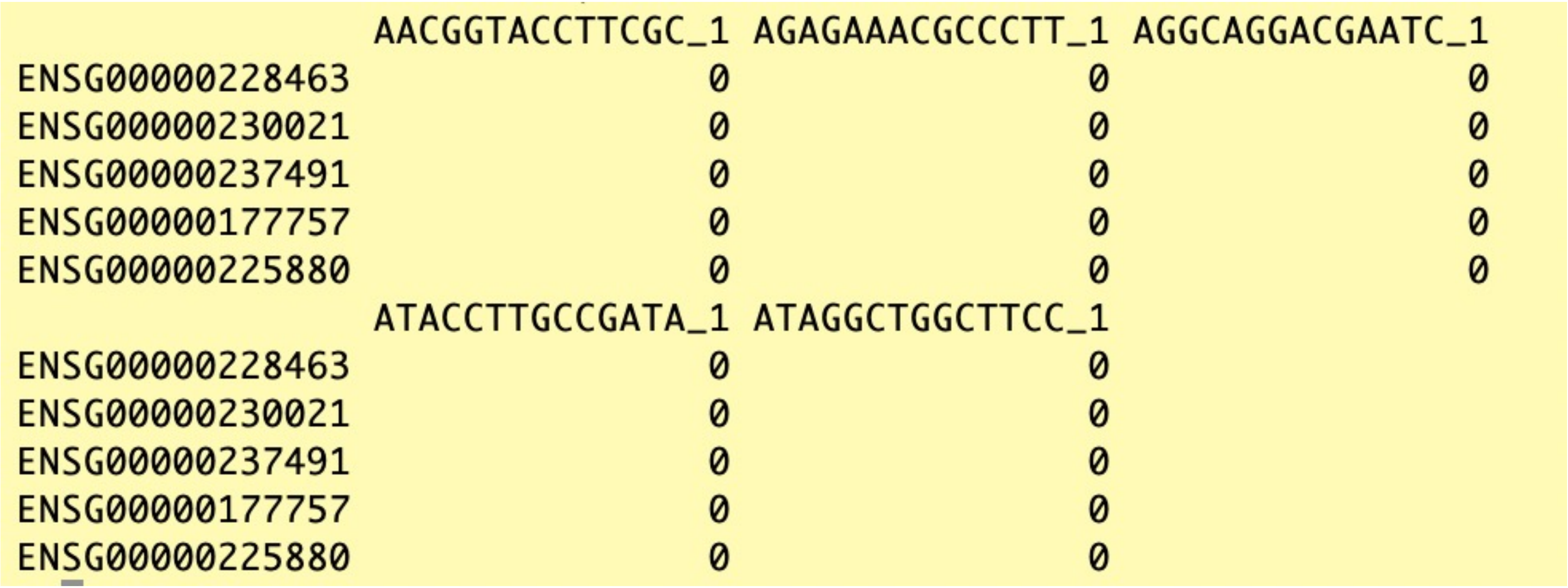


Figure source: Macosko et al. 2015, Potter SS. 2018

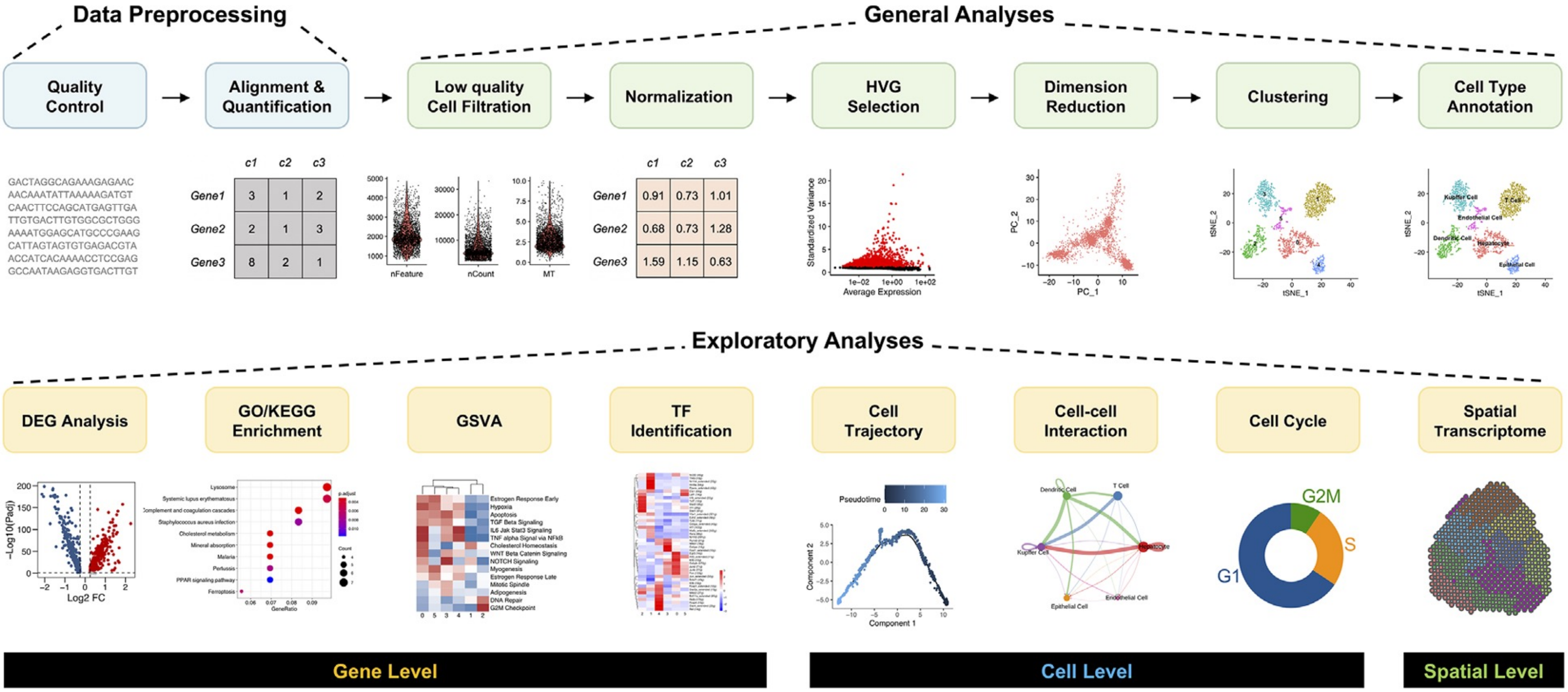
scRNA-seq data after processing

A matrix of read counts: rows are genes and columns are cells



```
                AACGGTACCTTCGC_1 AGAGAAACGCCCTT_1 AGGCAGGACGAATC_1
ENSG00000228463                0                0                0
ENSG00000230021                0                0                0
ENSG00000237491                0                0                0
ENSG00000177757                0                0                0
ENSG00000225880                0                0                0
                ATACCTTGCCGATA_1 ATAGGCTGGCTTCC_1
ENSG00000228463                0                0
ENSG00000230021                0                0
ENSG00000237491                0                0
ENSG00000177757                0                0
ENSG00000225880                0                0
```

Standard scRNA-seq data analysis pipeline



	c1	c2	c3
Gene1	3	1	2
Gene2	2	1	3
Gene3	8	2	1

	c1	c2	c3
Gene1	0.91	0.73	1.01
Gene2	0.68	0.73	1.28
Gene3	1.59	1.15	0.63

Figure source: Jovic, Dragomirka, et al. "Single-cell RNA sequencing technologies and applications: A brief overview." *Clinical and Translational Medicine* 12.3 (2022): e694.

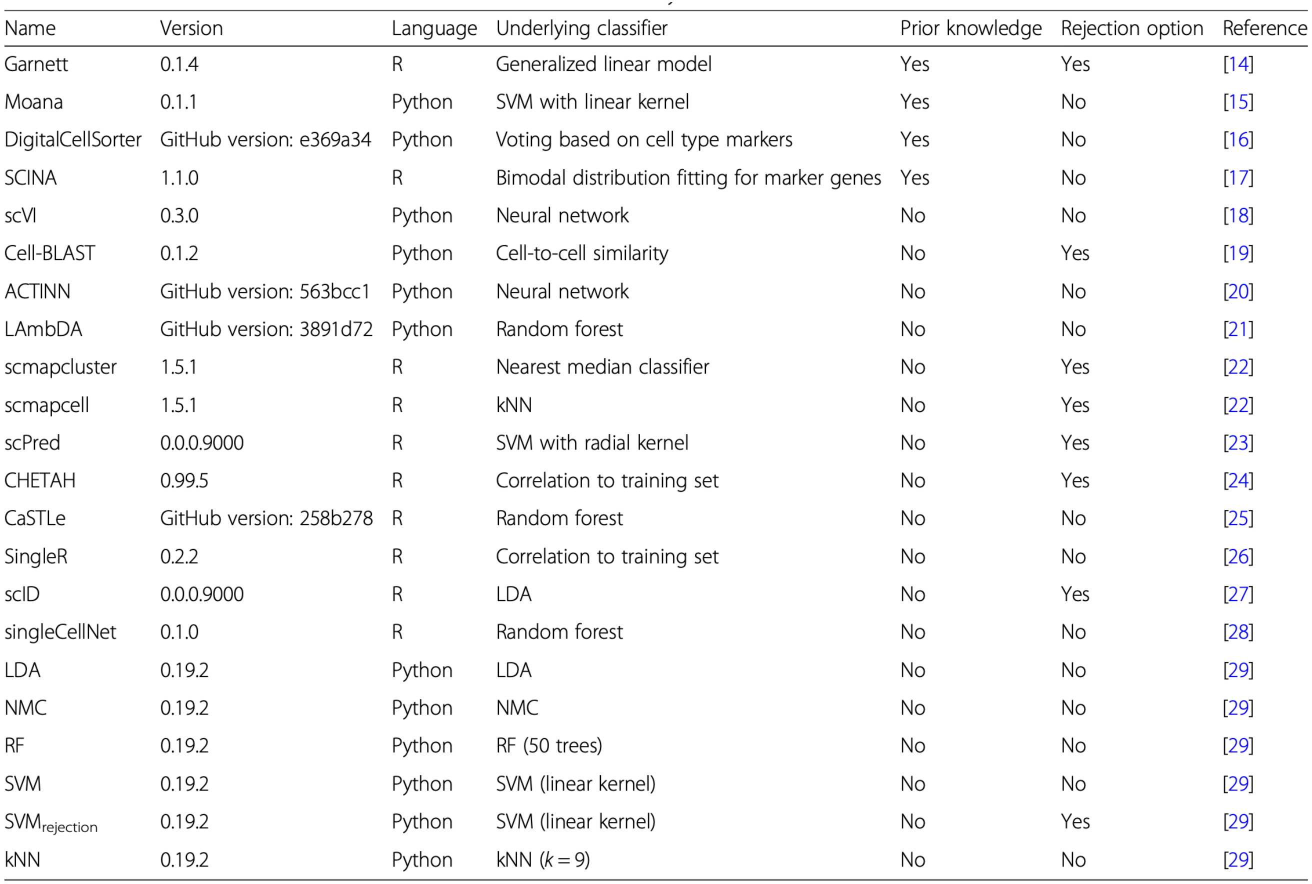
Challenges in identifying novel cells when annotating scRNA-seq data

1. Cell type annotation: one of the most important steps in scRNA-seq analysis
2. Traditional way of annotating cells: apply unsupervised clustering and label cell types based on the cluster-specific markers (Still widely used)
3. Supervised cell annotation methods have been developed to quickly and reproducibly assign cell labels. A comparison paper: Abdelaal et al. (2019, GB)

Pre-train a classifier using scRNA-seq training data with generic machine learning methods: SVM, LDA, RF, kNN, RF

- Scmap (Nature methods, 2018)
- CHETAH (NAR, 2019)
- singleR (Nat Immunol, 2019)

Challenges in identifying novel cells when annotating scRNA-seq data



Name	Version	Language	Underlying classifier	Prior knowledge	Rejection option	Reference
Garnett	0.1.4	R	Generalized linear model	Yes	Yes	[14]
Moana	0.1.1	Python	SVM with linear kernel	Yes	No	[15]
DigitalCellSorter	GitHub version: e369a34	Python	Voting based on cell type markers	Yes	No	[16]
SCINA	1.1.0	R	Bimodal distribution fitting for marker genes	Yes	No	[17]
scVI	0.3.0	Python	Neural network	No	No	[18]
Cell-BLAST	0.1.2	Python	Cell-to-cell similarity	No	Yes	[19]
ACTINN	GitHub version: 563bcc1	Python	Neural network	No	No	[20]
LAmbDA	GitHub version: 3891d72	Python	Random forest	No	No	[21]
scmapcluster	1.5.1	R	Nearest median classifier	No	Yes	[22]
scmapcell	1.5.1	R	kNN	No	Yes	[22]
scPred	0.0.0.9000	R	SVM with radial kernel	No	Yes	[23]
CHETAH	0.99.5	R	Correlation to training set	No	Yes	[24]
CaSTLe	GitHub version: 258b278	R	Random forest	No	No	[25]
SingleR	0.2.2	R	Correlation to training set	No	No	[26]
scID	0.0.0.9000	R	LDA	No	Yes	[27]
singleCellNet	0.1.0	R	Random forest	No	No	[28]
LDA	0.19.2	Python	LDA	No	No	[29]
NMC	0.19.2	Python	NMC	No	No	[29]
RF	0.19.2	Python	RF (50 trees)	No	No	[29]
SVM	0.19.2	Python	SVM (linear kernel)	No	No	[29]
$SVM_{rejection}$	0.19.2	Python	SVM (linear kernel)	No	Yes	[29]
kNN	0.19.2	Python	kNN ($k = 9$)	No	No	[29]

Challenges in identifying novel cells when annotating scRNA-seq data

1. Most of the conventional machine learning classification methods can only identify cell types that exist in the training data.

2. Existing methods generally rely on naïve approaches to identify novel cells:
   - Set a cutoff for correlation coefficients in scmap (default cutoff: 0.7)
   - Set a cutoff for confidence score of assignment in CHETAH (pc_thres = 0.2)
   - Set a cutoff for assigning probability in scPred (default value = 0.55)

3. Resulting in an excess number of unassigned cells (novel + uncertain cells)

Challenges in identifying novel cells when annotating scRNA-seq data

1. Neoplastic cells commonly exist in scRNA-seq data from cancer patients
2. One unique analytical challenge is distinguishing neoplastic cells (e.g., tumor cells) from nonneoplastic cells (e.g., immune cells, endothelial cells, and fibroblasts)
3. Cell sorting can be used as an experimental approach

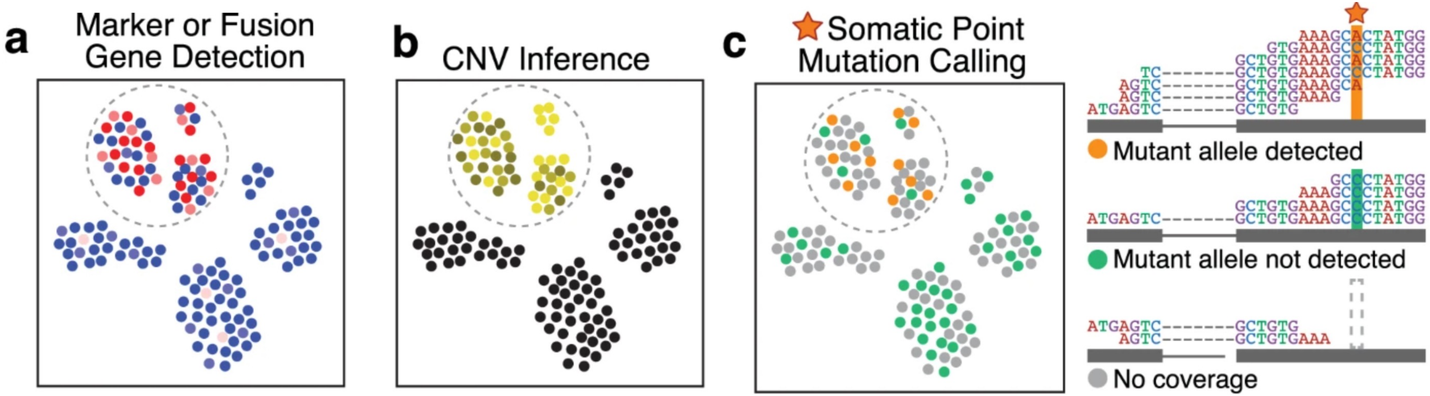
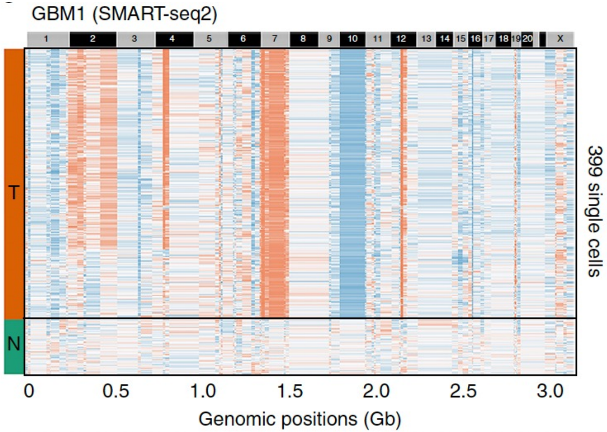


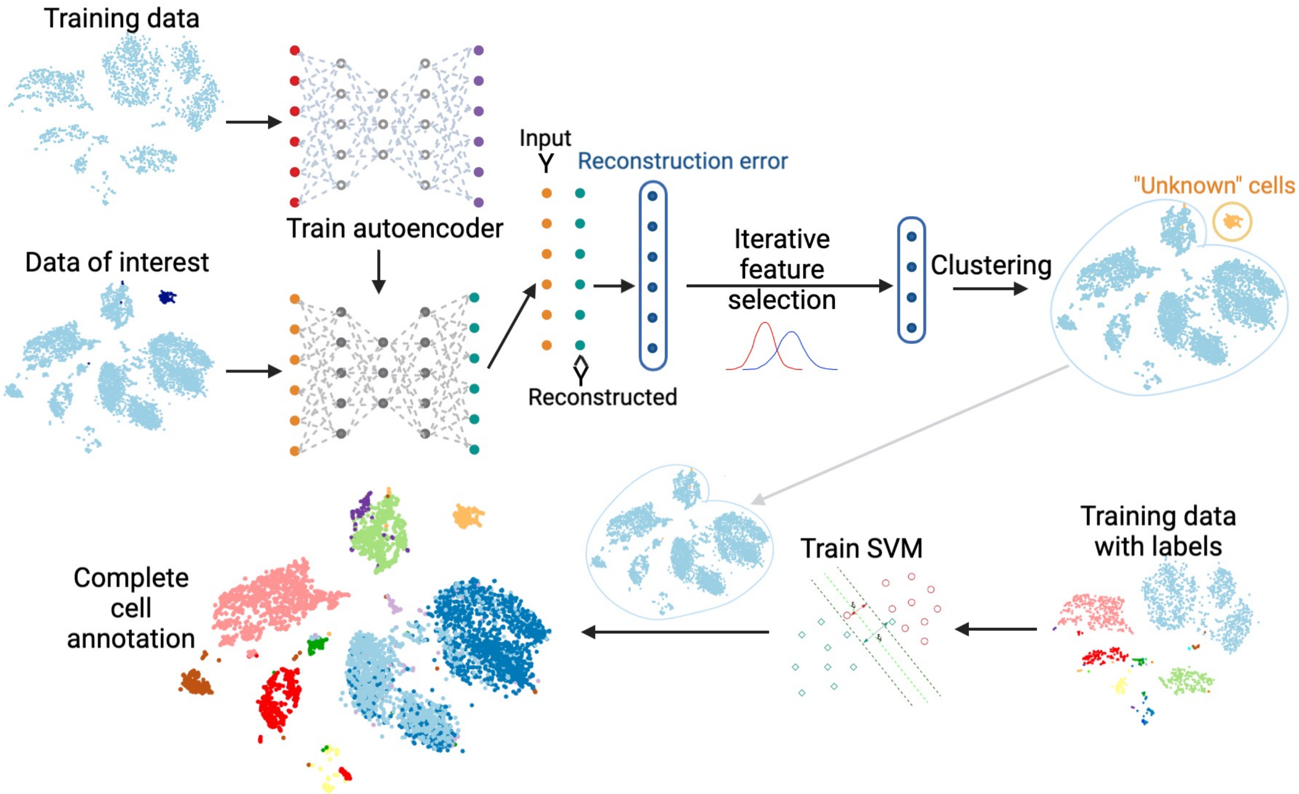
Figure source: Fan, et al. Experimental & Molecular Medicine 52.9 (2020): 1452-1465.

Challenges in identifying novel cells when annotating scRNA-seq data

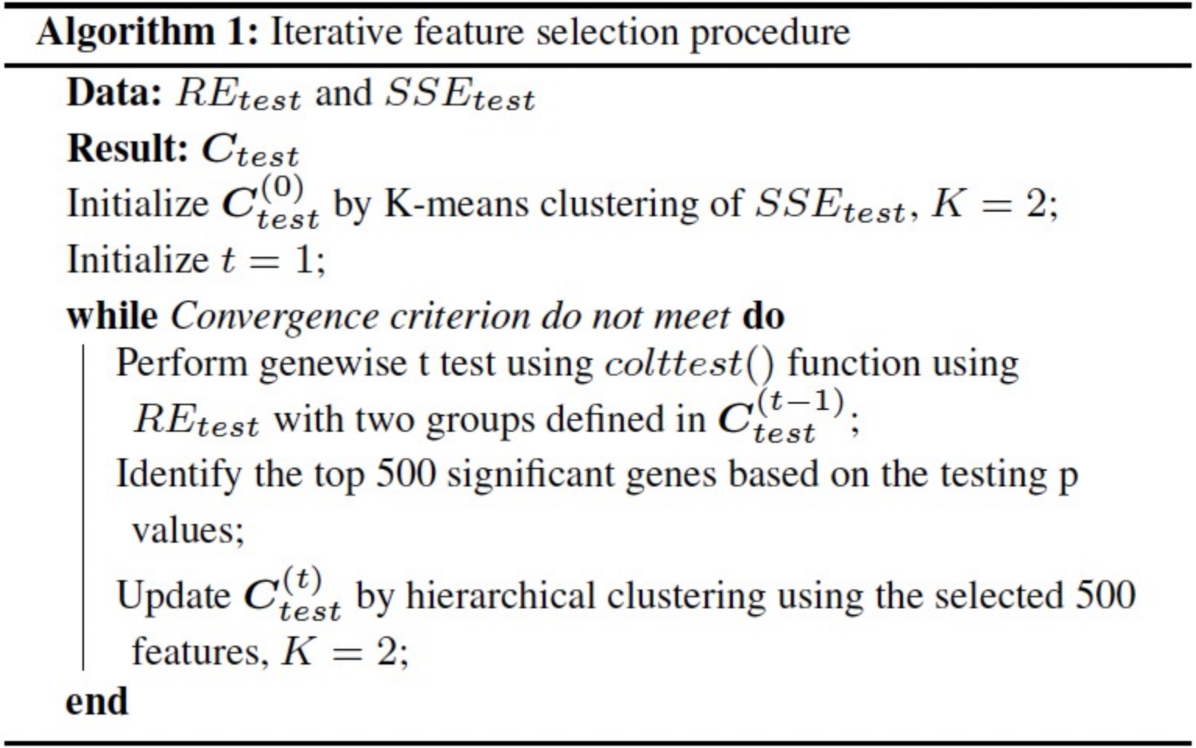
1. Computational methods have been developed to identify cells with extensive copy number variations
   - InferCNV (Science, 2014)
   - HoneyBadger (Genome Research, 2018)
   - CopyKAT (Nature Biotechnology, 2021)
2. These methods only works well when neoplastic cells have extensive copy number variations, but do not work when cells have small regions of aberrations or are diploid



Our proposal: a machine learning based method that does not rely on copy number variations



Our proposal: a machine learning based method that does not rely on copy number variations



```
Algorithm 1: Iterative feature selection procedure
Data: $RE_{test}$ and $SSE_{test}$
Result: $C_{test}$
Initialize $\boldsymbol{C}_{test}^{(0)}$ by K-means clustering of $SSE_{test}$, $K = 2$;
Initialize $t = 1$;
while Convergence criterion do not meet do
    Perform genewise t test using $colttest()$ function using
      $RE_{test}$ with two groups defined in $\boldsymbol{C}_{test}^{(t-1)}$;
    Identify the top 500 significant genes based on the testing p
      values;
    Update $\boldsymbol{C}_{test}^{(t)}$ by hierarchical clustering using the selected 500
      features, $K = 2$;
end
```

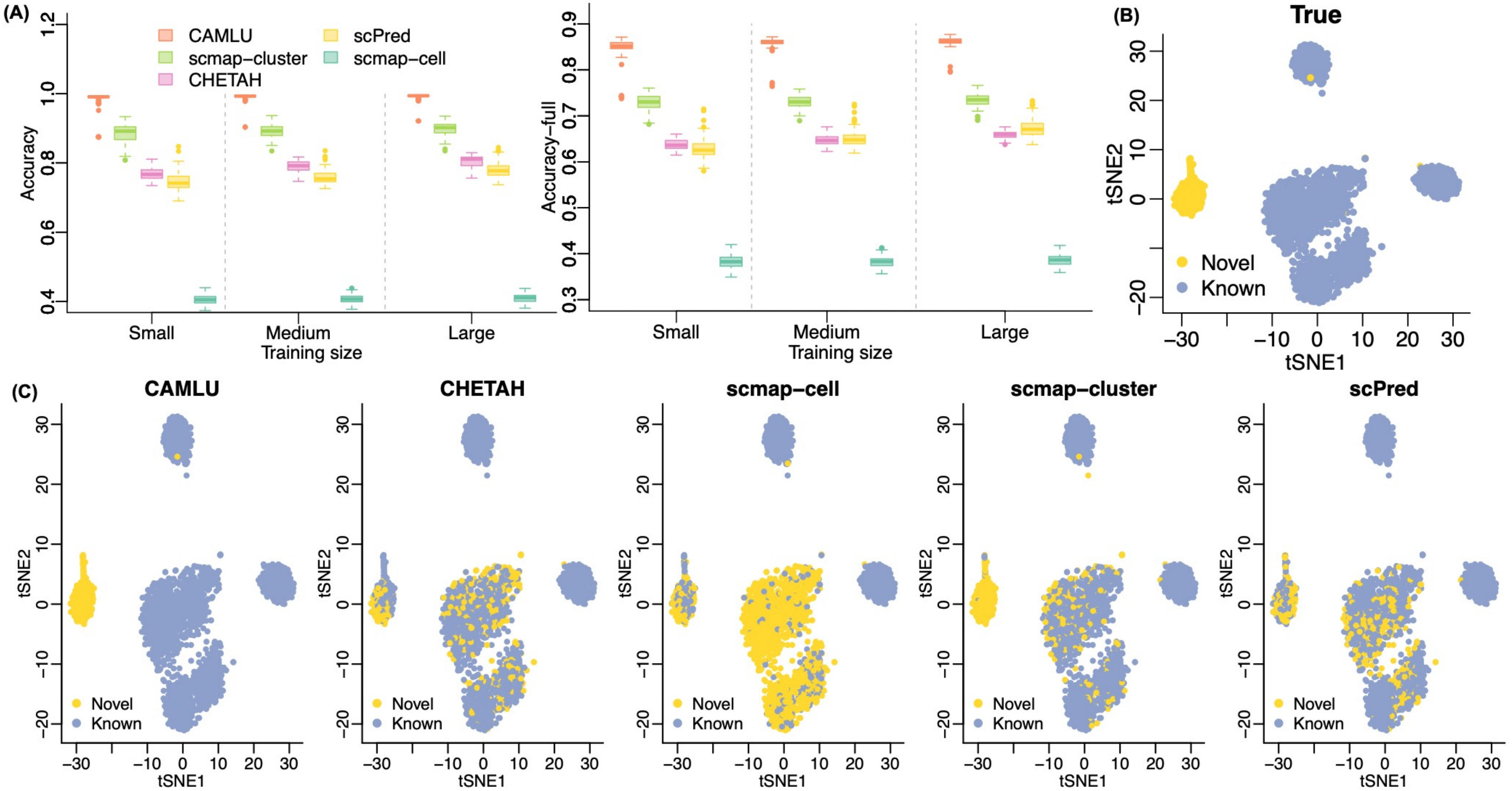
Designs of numeric experiments

- Three numerical experiments:
  - Peripheral blood mononuclear cells (PBMC, more than 60,000 sorted single cells), monocytes as the novel cell type
  - Draw training and testing data from the PBMC dataset excluding monocytes (n = 2400, 3100, 3800), add 300 monocytes in the test data

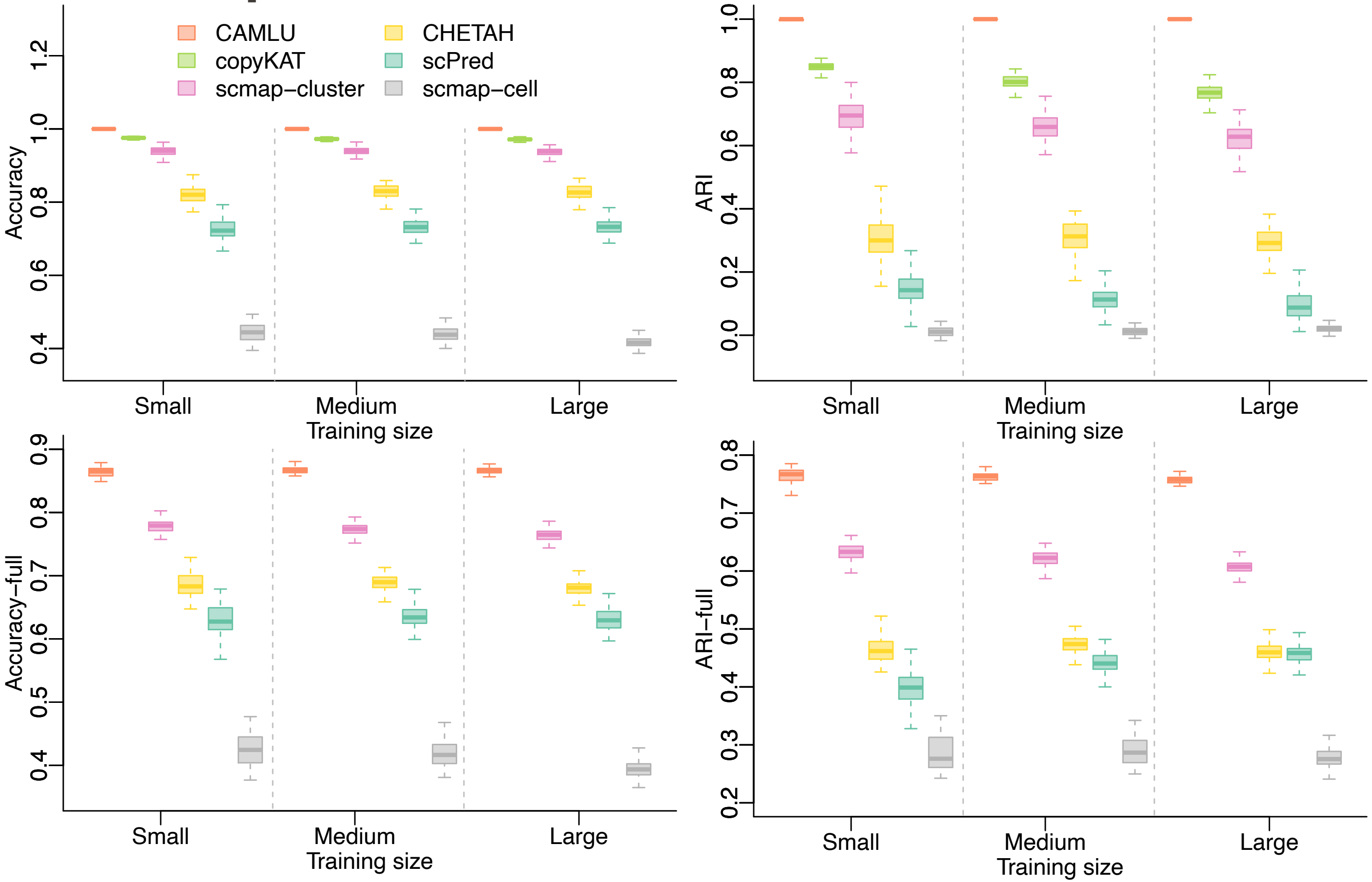
  - Peripheral blood mononuclear cells (PBMC, more than 60,000 sorted single cells) + head and neck cancer cell line (HNCC, 4632 cells)
  - Draw training and testing data from the PBMC dataset (n = 2400, 3100, 3800), add 300 cancer cells in the test data

  - Pancreas data (GSE85241, 2126 cells), mesenchymal cells as the novel cell type (80 cells)
- Comparing methods: CHETAH, scmap-cell, scmap-cluster, scPred, coypKAT (if cancer cells are involved)

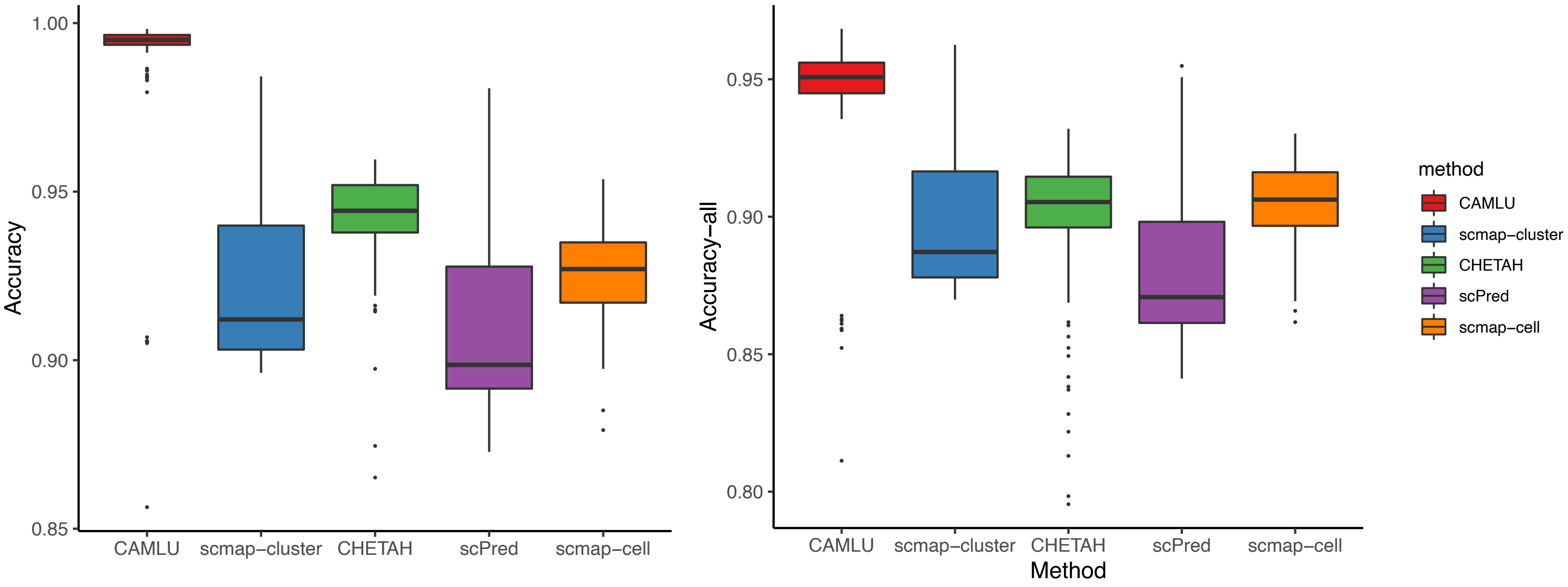
Numerical study with PBMC data



Numerical experiments with PBMC and cancer cell line data



Numerical study with Pancreas data

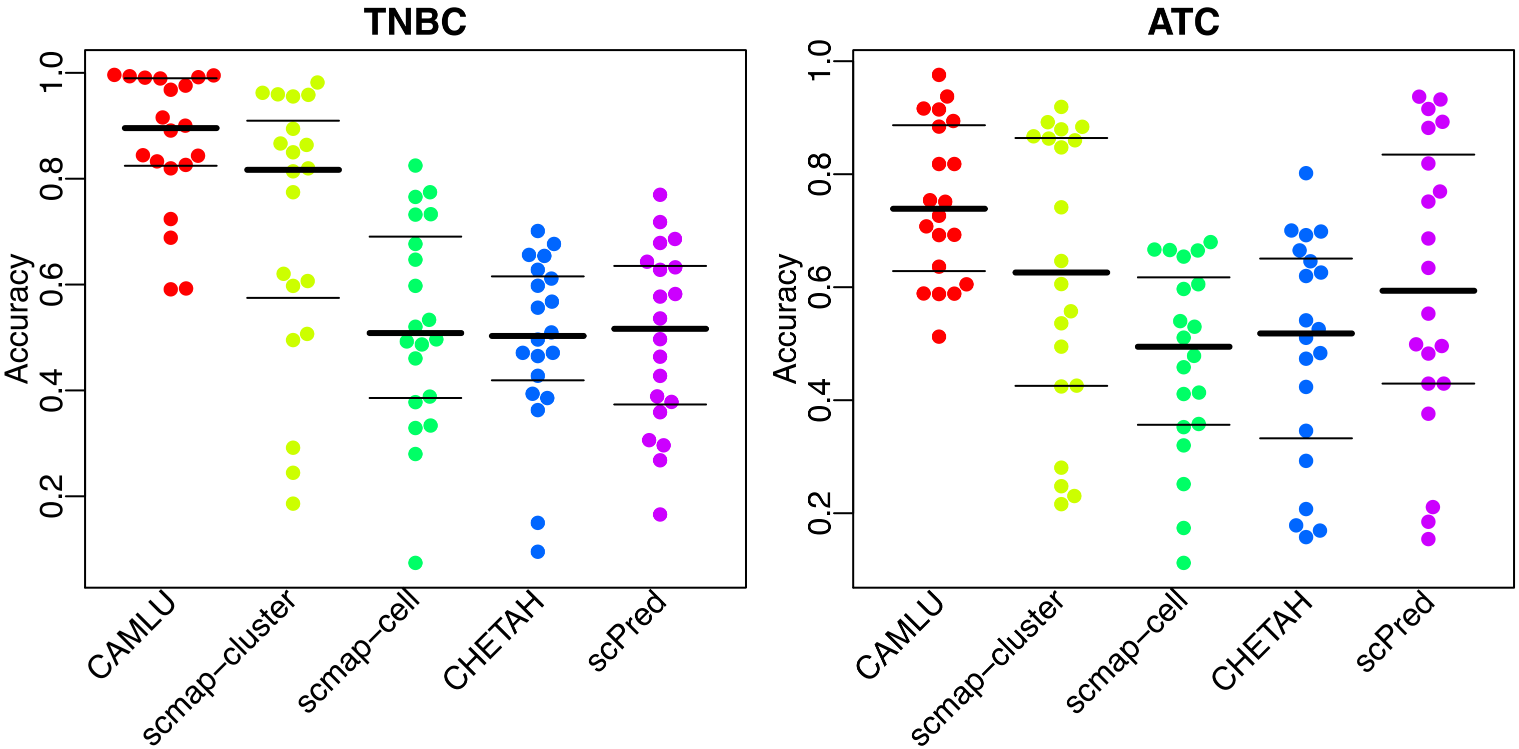


Application in two cancer datasets

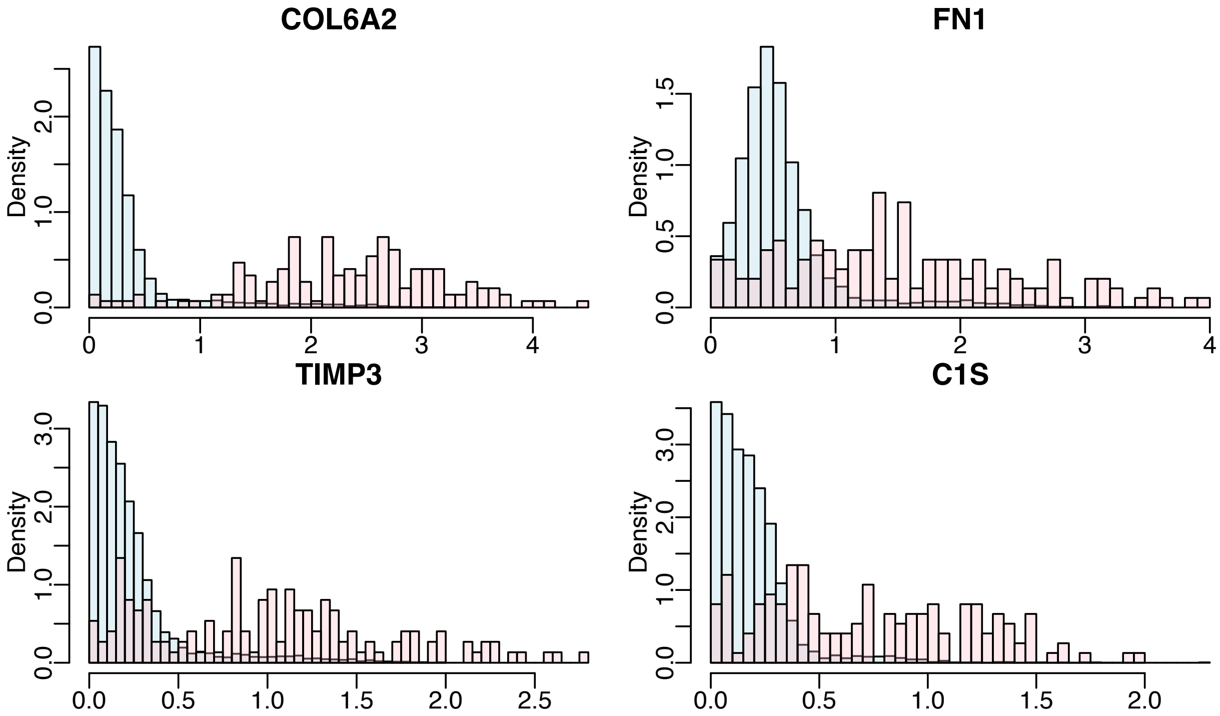
- scRNA-seq data with five triple-negative breast cancer (TNBC) patients
- scRNA-seq data with five anaplastic thyroid cancer (ATC) patients
- Both from Gao et al. (2021) and GSE148673

- Outside reference data for TNBC experiment: a scRNA-seq study with 26 primary tumors of three major breast cancer subtypes. The data from 10 TNBC patients were obtained as the reference.
- Wu et al. (2021) and GSE176078

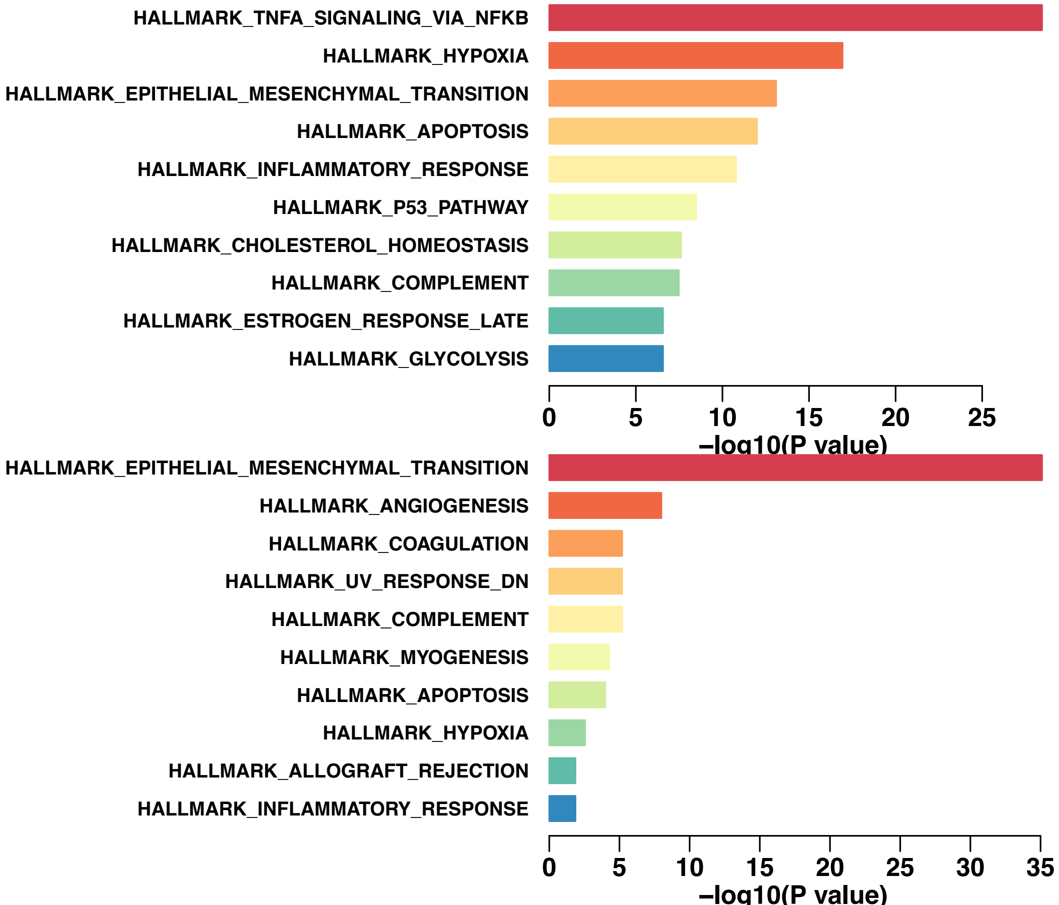
Application in two cancer datasets



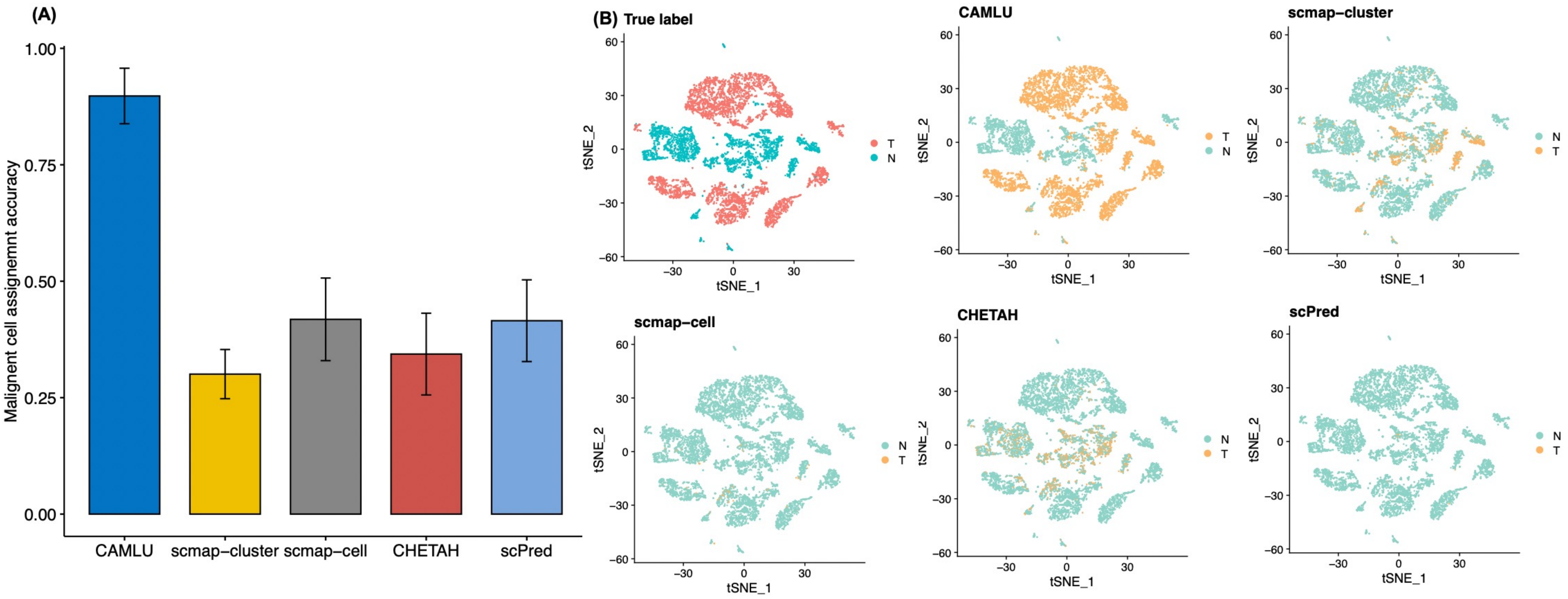
Application in two cancer datasets



Application in two cancer datasets



Analysis of TNBC data with external reference data



Unsolved challenges

- The proposed method may not work well when the novel cells are very similar to the known cells
- It is unclear if the method still works well if significant batch/subject effect exist in the data
- Will incorporating multiple reference panels improve classification accuracy?

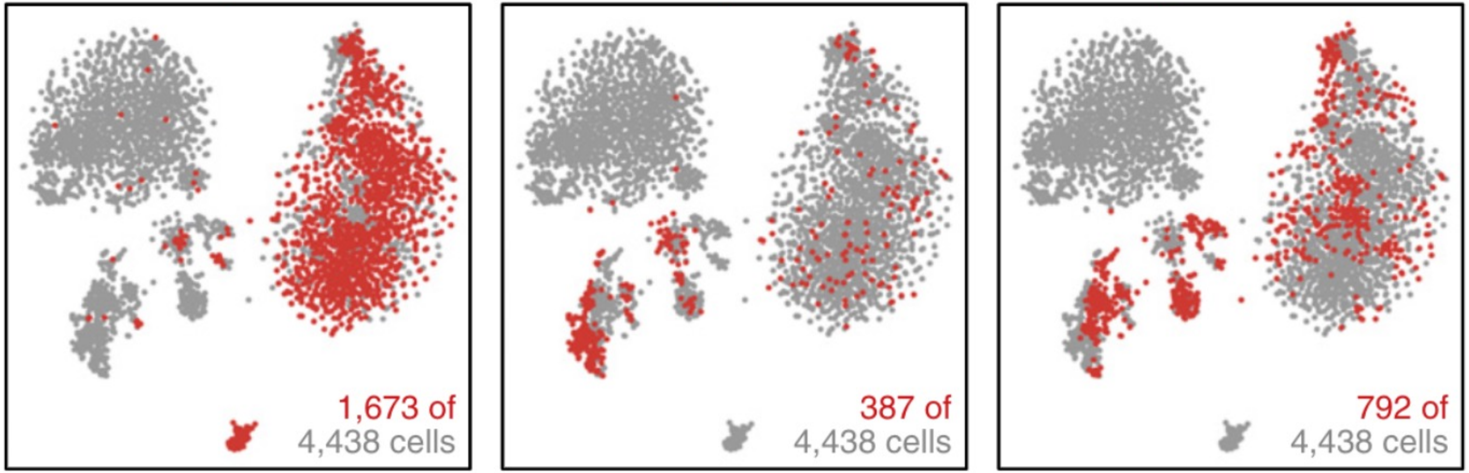
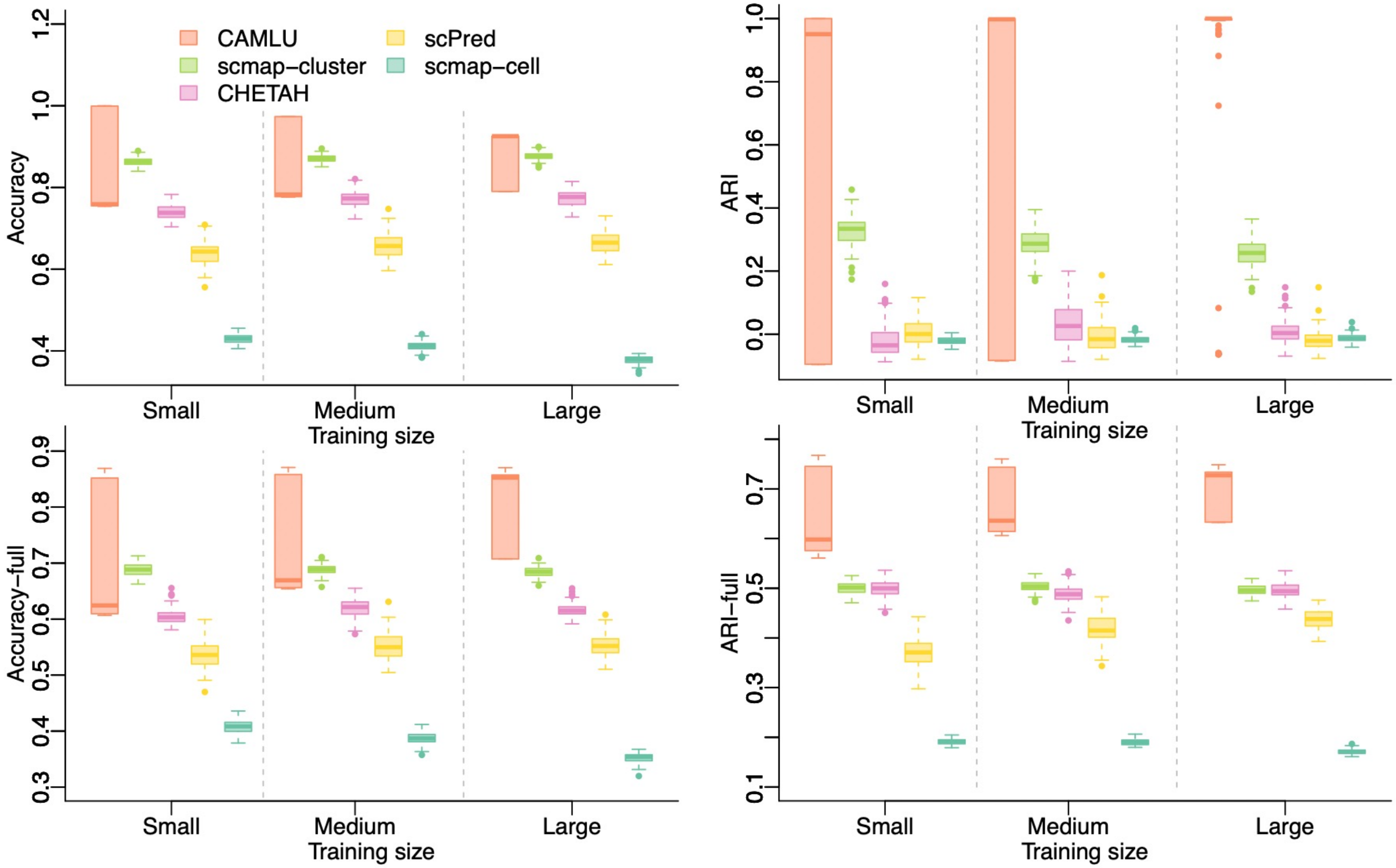


Figure source: van Galen, Peter, et al. "Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity." *Cell* 176.6 (2019): 1265-1281.

Unsolved challenges



Ongoing/future works

- Better identify neoplastic cells in certain cancer types by incorporating additional biological knowledge
- Including domain specific markers or pathway information to improve novel cell identification
- Explore this direction in larger population scale studies

https://ziyili20.github.io

Thank you!