Evaluation of some aspects in supervised cell type identification for single-cell RNA-seq

Wenjing Ma[1], Kenong Su[1], Hao Wu[1,2]

[1]Department of Computer Science, Emory University; [2]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University.

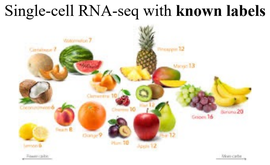
haowulab.org

Abstract

Cell type identification (cell typing) is one of the most important research questions in single-cell RNA sequencing (scRNA-seq) data analysis. Traditional annotation methods use unsupervised methods and marker gene expression to annotate clusters. However, it not only needs prior knowledge in biology but also its performance is affected by the number of cells and balancedness of cell type proportions in the dataset. With the accumulation of public scRNA-seq data, supervised cell typing methods have gained increasing popularity due to better accuracy, robustness, and computational performance.

Bulk RNA-seq Single-cell RNA-seq

Single-cell RNA-seq with **known labels**



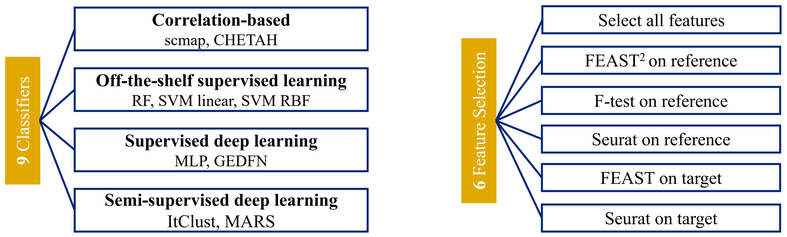
Despite all the advantages, the performance of the supervised cell typing methods heavily relies on several key factors: feature selection, supervised classifier, and most importantly, choice of the reference dataset. In our work[1], we have performed extensive real data analyses to systematically evaluate several strategies in supervised cell typing. Based on our analysis results, we provide guidelines as we suggest combining all individuals from available datasets to construct the reference dataset and use Multi-Layer Perceptron (MLP) as the classifier, along with F-test as the feature selection method.

Code access: https://github.com/marvinquiet/RefConstruction_supervisedCelltyping

Benchmarking

We benchmarked following factors:

- **Classifiers and feature selection methods**

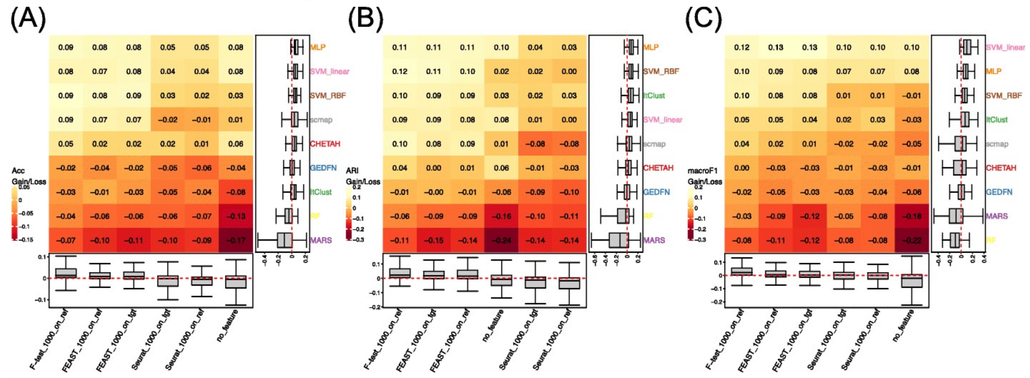


- **No. of cells and No. of cell types in the reference dataset**
- **Data preprocessing**
  - ❖ Batch effect removal: Harmony, fastMNN
  - ❖ Data imputation: MAGIC, SAVER, scVI
- **Discrepancies between reference and target datasets**
  - ❖ Individual difference: when reference and target data are from different individuals. Thus, the discrepancy only comes from biological variations.
  - ❖ Condition difference: when reference and target data are from different conditions, such as protocol difference (10X Chromium vs. Smart-Seq2), sample collection difference (frozen vs. fresh tissues), biological difference (e.g., different brain regions), and clinical difference (e.g., different disease status), etc.
- **Pooling and purifying samples**
- **Discovering new cell types**

In total, we benchmarked 29 experiments derived from 10 real datasets (including mouse brain, human PBMC and human pancreas). We chose three evaluation metrics: Accuracy (Acc), Adjusted Rand Index (ARI) and Macro F1.

Results

Prediction performance gains/losses with different combinations of classifiers and feature selection methods

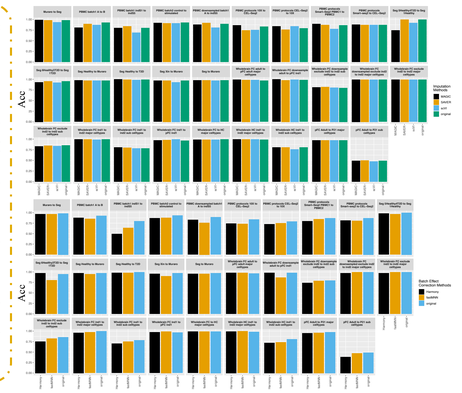


For the 9x6 performance table X in each experiment, we first compute a baseline: $\bar{X}$ and use $X - \bar{X}$ to remove the experimental effect. We then calculate $\overline{X - \bar{X}}$ and sort the values for each row and column so that top left corner shows the largest performance gain.

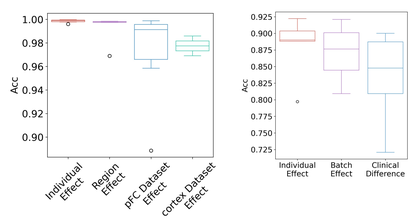
As for marginal boxplots, same procedure has been done as we remove feature selection effect for calculating gains/losses of using different classifiers and same for feature selection.

Performance comparison between original datasets and preprocessed datasets

We conducted the aforementioned imputation methods (upper panel) on both reference and target and removed batch effect (lower panel) between reference and target datasets. No single imputation/batch effect removal method outperforms others.



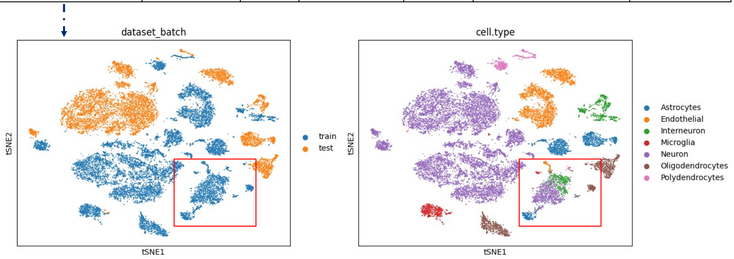
Different discrepancies between reference and target



As shown in the left panel, dataset effect has the most impact on prediction. In addition, in the right panel, clinical difference affects most.

Pooling effect analysis

	Before (mean Acc)	After (Acc)	Before (mean ARI)	After (ARI)	Before (mean Macro F1)	After (Macro F1)
Mouse brain region effect	0.993	**0.996**	0.995	**0.997**	0.915	**0.959**
Mouse brain dataset effect (pFC)	0.971	**0.988**	0.947	**0.967**	0.892	**0.904**
Mouse brain dataset effect (cortex)	0.977	**0.982**	0.955	**0.965**	0.918	**0.927**
Mouse brain dataset effect (combine pFC and cortex)	-	**0.986**	-	**0.963**	-	**0.933**



The table shows performance improvements (bold font) when pooling individuals or datasets together to predict. Although the two datasets do not mix well (as shown in the figure), the Accuracy, ARI and Macro F1 increase compared to using individual dataset.

Guidelines

- **Classifier and feature selection is important**: in our analysis, MLP + F-test achieves the best performance;
- **Condition effect exists but tolerable**: condition effect between reference and target dataset can affect performance, but when no matching condition exists, dataset from the same tissue can still be used as reference dataset;
- **Pooling samples from different datasets improve the prediction**: we suggest combining all individuals/samples together as reference to average out the individual or condition variation;
- **Purifying the dataset by removing noisy cells does not significantly improve the results**;
- **Sub-cell types classification is tricky**: when lacking consistency of sub-cell types between datasets, we recommend first using supervised cell typing methods to classify major cell types and then unsupervised clustering as a second step.

References

[1]Ma, W., Su, K. and Wu, H., 2021. Evaluation of some aspects in supervised cell type identification for single-cell RNA-seq: classifier, feature selection, and reference construction. Genome biology, 22(1), pp.1-23.

[2]Su, K., Yu, T. and Wu, H., 2021. Accurate feature selection improves single-cell RNA-seq cell clustering. Briefings in Bioinformatics, 22(5).