CMML ICA2 - Task 8

Supporting Materials

Task 8 2074

Supplementary Methods

Code Availability

Code is available at https://github.com/guobiao-ye/Cell_Annotation_Tools_Benchmark.git.

Data Acquisition

Publicly available scRNA-seq datasets, CellBench 10X (GSM3618014) and CellBench CEL-Seq2 (GSM3618022-GSM3618024), served as primary data sources (Abdelaal et al., 2019; Tian et al., 2019). The labels from the dataset were used as the ground truth annotations.

Dataset Partitioning and Reference Simulation

To assess method performance across varying reference data availabilities, datasets were systematically partitioned. For intra-dataset evaluations (CEL-Seq2 cohort), cells were allocated into reference and query sets via stratified sampling based on ground truth labels, generating reference populations of 75%, 25%, 50%, 12.5%, 5%, 2.5% and 1.25% of total cells, with reproducibility ensured via fixed random seeds. For experiments probing batch effects and larger data volumes, 10X and CEL-Seq2 datasets were merged and similarly partitioned. All partitioning yielded standardized count matrices and label files for reference and query sets.

Implementation and Execution of Annotation Algorithms

Three annotation algorithms were benchmarked: *SingleR* (v2.4.1), *scmap* (v1.24.0), and *scANVI* (v1.3.1), using R (v4.3.2) and Python (v3.12) implementations.

SingleR and scmap-cell. Data were structured in *SingleCellExperiment* objects. Preprocessing included gene filtering (mean counts < 0.01) and log-normalization **(**converting to Counts Per Million followed by a log1p transformation**)**. *SingleR* used default parameters, typically involving correlation against reference profiles using differentially expressed genes for weighting, after gene list harmonization. For *scmap*, reference feature selection (retaining 500 features) preceded index construction (k-nearest neighbors, k=10) and query cell projection. Processed query data objects were saved for UMAP visualization. The scripts were implemented and run in the *R* environment.

scANVI. Data were managed as *AnnData* objects (*scanpy* and *scvi-tools* ecosystem). Preprocessing involved library size normalization (target sum 1e4) and log1p transformation. Highly variable genes (2000 top) from the reference defined a common feature space. An *scVI* model (variational autoencoder; 2 layers, 30 latent dimensions, negative binomial likelihood) was trained on the aligned reference (max 200 epochs, early stopping), then converted and fine-tuned as an scANVI model using reference labels (max 100 epochs, early stopping). Predictions were generated on aligned query data, using GPU acceleration. Processed query data objects were also saved. The scripts were implemented and run in the *python* environment.

Prediction outputs were standardized, mapping cell identifiers to true and predicted labels. Execution times, peak CPU RAM and GPU VRAM (for scANVI) usage were further tested and systematically recorded.

Benchmark Analysis and Performance Visualization

A unified *R*-based analytical pipeline performed comparative assessment. Prediction outputs were aggregated, focusing on cells common to all methods for equitable comparison. Performance was evaluated using overall accuracy, Cohen's Kappa, and per-class precision, recall, and F1-score (via established *R* packages). For UMAP visualizations, coordinates were derived from query log-counts using a standard *R ump* implementation. Visual outputs (UMAPs, confusion matrices, performance bar plots) were generated using *ggplot2*. Furthermore, time and memory footprints were visualized as line plots and/or stacked bar charts to illustrate resource consumption across different dataset proportions. Comprehensive metric reports were produced per run.

Reference

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J. T., & Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome biology*, *20*(1), 194.

Tian, L., Dong, X., Freytag, S., Lê Cao, K. A., Su, S., JalalAbadi, A., Amann-Zalcenstein, D., Weber, T. S., Seidi, A., Jabbari, J. S., Naik, S. H., & Ritchie, M. E. (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature methods*, *16*(6), 479–487.

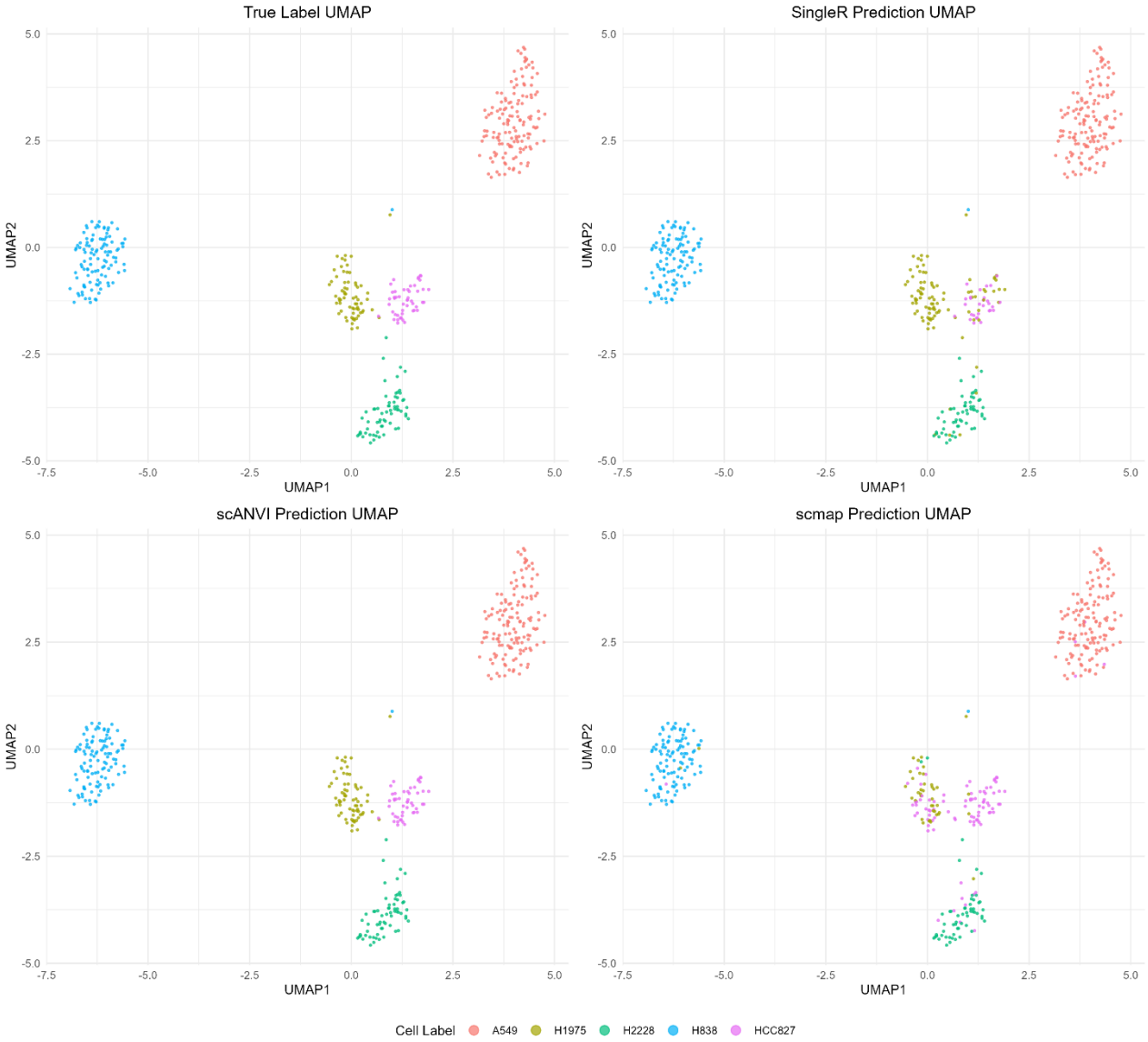
Reflection

This miniproject was a practical application of CMML3 concepts, focusing on reference mapping for cell type annotation and tackling batch effects in scRNA-seq. Benchmarking *SingleR*, *scmap*, and *scANVI* allowed me to directly compare how correlation-based, projection-based, and VAE-based methods (like *scANVI*) perform annotation. Performance variations across reference scales underscored the need to understand algorithmic designs: *scANVI*'s VAE excelled with larger, heterogeneous data by learning batch-corrected latent spaces, while *SingleR*'s rank-based approach proved robust with extremely scarce, homogenous references. Future work could involve advanced simulations (e.g., using *scDesign* principles) for deeper sensitivity analysis, and exploring Foundation Models for annotation. CMML3 modules on statistical vs. deep learning methods for scRNA-seq (simulation, reference mapping, batch correction), VAE mechanisms, and Foundation Model principles were crucial. They provide insights into the implementation of benchmarks and interpretation of results based on software features. This project honed my autonomy in designing, executing, and critically evaluating bioinformatics comparisons, skills vital for future research where robust data integration and advanced algorithmic modeling are essential.

Supplementary Table and Figures

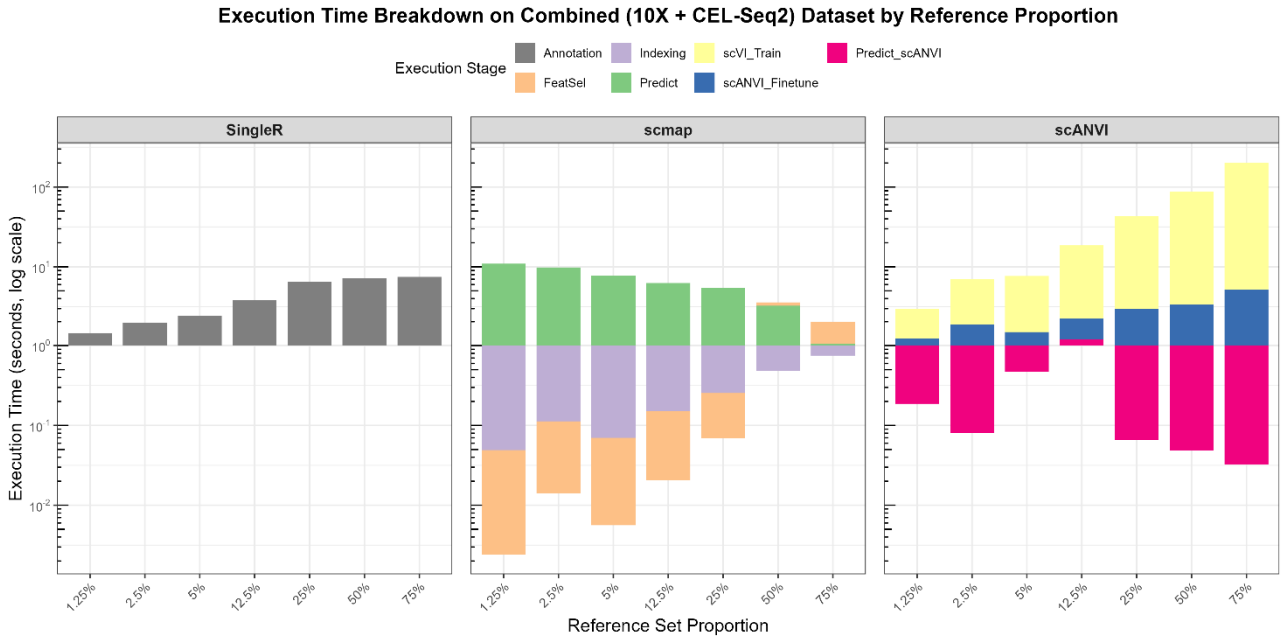
Ref. Proportion	Software	Total Time (s)	Peak CPU RAM (GB)	Peak GPU VRAM (GB)
75.00%	SingleR	7.48	3.17	N/A
(No. ref=3280)	scANVI	44.31	12.56	6.73
	scmap	3.71	2.95	N/A
50.00%	SingleR	7.14	3.05	N/A
(No. ref=2186)	scANVI	29.66	9.34	4.66
	scmap	4.85	2.76	N/A
25.00%	SingleR	6.53	2.25	N/A
(No. ref=1093)	scANVI	17.55	6.14	3.26
	scmap	6.01	2.08	N/A
12.50%	SingleR	3.86	1.55	N/A
(No. ref=547)	scANVI	11.32	4.19	2.14
	scmap	6.55	1.64	N/A
5.00%	SingleR	2.46	1.11	N/A
(No. ref=219)	scANVI	7.13	2.75	1.45
	scmap	7.95	1.39	N/A
2.50%	SingleR	1.94	0.84	N/A
(No. ref=109)	scANVI	5.73	2.14	1.11
	scmap	10.04	1.25	N/A
1.25%	SingleR	1.44	0.69	N/A
(No. ref=55)	scANVI	3.83	1.51	0.86
	scmap	11.17	1.32	N/A

Supplementary Table 1 | Computational Resource Consumption.



Supplementary Figure 1 | UMAP visualizations of cell type annotation on the CEL-Seq2 dataset using a 25% reference proportion.

Uniform Manifold Approximation and Projection (UMAP) plots of the CEL-Seq2 query cells. Top left: Cells colored by ground truth labels. Top right: Cells colored by SingleR predictions. Bottom left: Cells colored by scANVI predictions. Bottom right: Cells colored by scmap-cell predictions. All methods were provided with a reference set comprising 25% of the total CEL-Seq2 cells.



Supplementary Figure 2 | Execution time breakdown for annotation methods on the combined dataset across reference proportions.

Stacked bar charts illustrating the proportion of total execution time attributed to distinct computational stages for SingleR, scmap-cell, and scANVI when applied to the combined 10x Chromium and CEL-Seq2 dataset. Reference set proportions range from 1.25% to 75%. For SingleR, "Annotation" represents the primary computational step. For scmap-cell, stages include "*FeatSel*" (Feature Selection), "*Indexing*," and "*Predict*." For scANVI, stages are "*scVI_Train*," "*scANVI_Finetune*," and "*Predict_scANVI*."