

CMML ICA2 - Main Report

Topic: Automatic cell type annotation

Task 8 2074

Pragmatic Robustness in Automated Cell Annotation: Navigating Reference Scarcity, Batch Heterogeneity, and Computational Constraints with SingleR, scmap, and scANVI

Abstract

Automated cell annotation is vital for scRNA-seq, yet tool performance varies with real-world data. We benchmarked SingleR, scmap, and scANVI across reference scales, batch effects, and runtimes. scANVI excelled with larger, heterogeneous references but incurred higher training costs. SingleR showed striking accuracy and speed with extremely scarce homogenous references. scmap benefited from reference size but faced prediction scalability issues. Findings guide pragmatic tool selection by balancing algorithmic strengths against data/computational realities.

Main Text

The transformative power of single-cell RNA-sequencing (scRNA-seq) hinges on accurate cell type annotation, a process increasingly reliant on automated methods to handle escalating data scales and complexity (Cheng et al., 2023). However, the "ideal" scenario of abundant, perfectly matched, and computationally tractable reference data rarely mirrors experimental reality. Researchers frequently encounter limitations such as scarce reference exemplars, pervasive batch effects when integrating diverse datasets, and finite computational resources (Pasquini et al., 2021). Navigating these pragmatic constraints necessitates a deep understanding of how different annotation paradigms, rooted in distinct algorithmic principles, cope under pressure. We therefore dissected the performance and efficiency of three archetypal annotators—correlation-based SingleR, projection-based scmap-cell, and deep generative model-based scANVI—across a deliberately challenging spectrum of reference data availability, inter-dataset heterogeneity, and computational demand, interpreting their performance in light of their core mechanisms (Table 1).

Our investigation first probed the fundamental limit of reference scale through intra-dataset partitioning within a controlled, homogenous CEL-Seq2 dataset (5 labels, 570 cells). By systematically varying the proportion of cells used for reference versus query, we simulated scenarios of differing reference data availability. While all methods performed well with an ample reference subset (25% of cells designated as reference), a critical divergence, driven by algorithmic sensitivities, emerged under extreme scarcity (Fig. 1a; Supplementary Fig. 1). When the reference was winnowed to a mere 2.5%-5% ($n=14-28$ cells), SingleR displayed remarkable robustness (Accuracy > 0.98) (Fig. 1a, b). This resilience likely stems from its core Spearman rank correlation approach, which emphasizes relative gene expression rankings rather than absolute magnitudes. Such rank-based measures can retain discriminatory power even with few noisy exemplars per cell type, as the characteristic "shape" of a cell type's transcriptome might be preserved. In stark contrast, scANVI (Accuracy=0.624-0.811) and scmap (Accuracy=0.309-0.601) faltered significantly. scANVI, a variational autoencoder (VAE), likely requires a sufficient number of diverse data points per labeled group to accurately model the data distribution and learn a meaningful, disentangled latent space. With too few reference cells, the VAE may overfit or fail to capture the true underlying manifold structure, leading to poor generalization. Similarly, scmap-cell, which projects query cells onto a reference indexed by k -nearest neighbors or centroids using cosine similarity after feature selection, may suffer from unstable feature selection and an insufficiently populated reference manifold when reference exemplars are exceedingly rare, making projections unreliable.

Building upon this, we explored the interplay between reference scale and batch heterogeneity by creating a combined 10x Chromium (5 labels, 3803 cells) and CEL-Seq2 dataset, which simulated the common challenges in broader and multi-source reference maps. With a substantial 75% combined reference, scANVI's VAE architecture, designed for probabilistic harmonization of datasets, demonstrated its strength by achieving perfect accuracy and effectively mitigating batch effects, albeit with the highest training cost (44.31s, 12.56 CPU RAM and 6.73 GPU VRAM) (Fig. 1c; Supplementary Tabel 1). This aligns with its intended use for integrating multiple datasets by learning a shared representation that removes unwanted variation. SingleR (Accuracy=0.978), by correlating against reference profiles that inherently average out some noise, and scmap (Accuracy=0.999), by projecting onto a now richer and more diverse reference manifold, also performed exceptionally. More revealingly, even when this heterogeneous reference was proportionally small (e.g., 2.5%, n=109 cells), scANVI maintained high accuracy (0.980). This suggests that its ability to model and correct for batch variation given sufficient *absolute* cell numbers (even if proportionally few from a large combined pool) is a key advantage. SingleR's performance remained robust, likely because its pairwise correlations are somewhat inherently resilient to global shifts if marker gene rankings are preserved. scmap's accuracy, while good, was slightly lower than scANVI and SingleR in this specific low-proportion heterogeneous scenario, potentially because its feature selection (e.g., based on median expression) might be more sensitive to batch-driven expression differences when reference diversity is high but exemplar count per batch within each cell type is low. Computationally, SingleR's direct correlation approach remained exceptionally resource efficient, while scANVI's VAE training, even if shortened with fewer cells, represented a comparatively larger fixed cost for model building (Fig. 1d, e; Supplementary Fig. 2). Besides, scmap's projection step exhibited a notable increase in prediction runtime with the larger query set, a consequence of its k-NN search or distance calculations scaling with query size, posing a computational scalability concern for very large target datasets (Fig. 1d; Supplementary Fig. 2).

The generalization to unseen batch-distinct data (training on 10x, testing on CEL-Seq2) showed high accuracy (>98%) for all tools. This indicates that when a reference is large and sufficiently captures cell type signatures, even if from a different batch, the core biological signals can be strong enough for diverse algorithms (correlation, projection, generative modeling) to succeed. scANVI's inherent design for domain adaptation likely contributes here, while SingleR and scmap benefit if the distinguishing features remain relatively consistent in their rankings or selected space.

Collectively, our findings, by linking performance to algorithmic principles, delineate a pragmatic decision framework. For rapid annotation from extremely limited, relatively homogenous references, SingleR's efficient, rank-based correlation offers unparalleled utility. When integrating larger, diverse, and batch-confounded references, scANVI's deep generative modeling provides superior accuracy and integration, justifying its higher training cost if a critical mass of reference cells is available. scmap, relying on feature selection and projection, is effective with well-populated references but requires attention to prediction scalability. This study underscores that optimal tool selection demands a nuanced appreciation of how algorithmic foundations interact with the practical realities of reference data scale, heterogeneity, and computational constraints, guiding researchers toward more robust and efficient automated cell type identification.

Table and Figure

Dataset Feature	CellBench 10X	CellBench CEL-Seq2	
GEO Accession	GSM3618014	GSM3618022, GSM3618023, GSM3618024	
Platform	10X Chromium	CEL-Seq2	
Cell Lines	A549, H1975, H2228, H838, HCC827	A549, H1975, H2228, H838, HCC827	
ddNo. Cells	3,803	570	
No. Labels	5	5	
Software Feature	SingleR	Scmap (scmap-cell)	scANVI
Method Type	Correlation-based	Distance-based	Deep learning (variational autoencoder)
Version	2.4.1	1.24.0	1.3.1
Input	scRNA-seq expression matrix, reference dataset	scRNA-seq expression matrix, reference dataset	scRNA-seq expression matrix, optional batch info
Annotation Approach	Cell-level, gene correlation	Cell-level or cluster-level, nearest neighbor	Cell-level, latent space embedding
Reference Requirement	Bulk or single-cell reference	Single-cell reference	Single-cell reference, optional labels
Output	Cell type labels, confidence scores	Cell type labels, similarity scores	Cell type labels, latent representations
Batch Correction	Limited, requires pre-processing	Limited, requires pre-processing	Integrated batch correction
Speed	Fast	Fast	Slower (GPU-dependent)
Scalability	High, efficient for large datasets	Good for moderate datasets	Moderate, computationally intensive
Feature Selection	Automatic (highly variable genes)	User-defined or automatic	Automatic (learned embeddings)
Unlabeled Data Handling	Predicts labels only	Predicts labels, rejects low-confidence	Predicts labels, integrates unlabeled data
Robustness	Robust to noise, sensitive to reference	Sensitive to reference quality	Robust, handles batch effects well
Implementation	R (Bioconductor)	R (Bioconductor)	Python (scvi-tools)
Use Case	General cell type annotation	Fast annotation, large datasets	Complex datasets, batch integration
Reference	Aran et al., 2019	Kiselev et al., 2018	Xu et al., 2021

Table 1 | Overview of Datasets and Benchmarked Annotation Software.

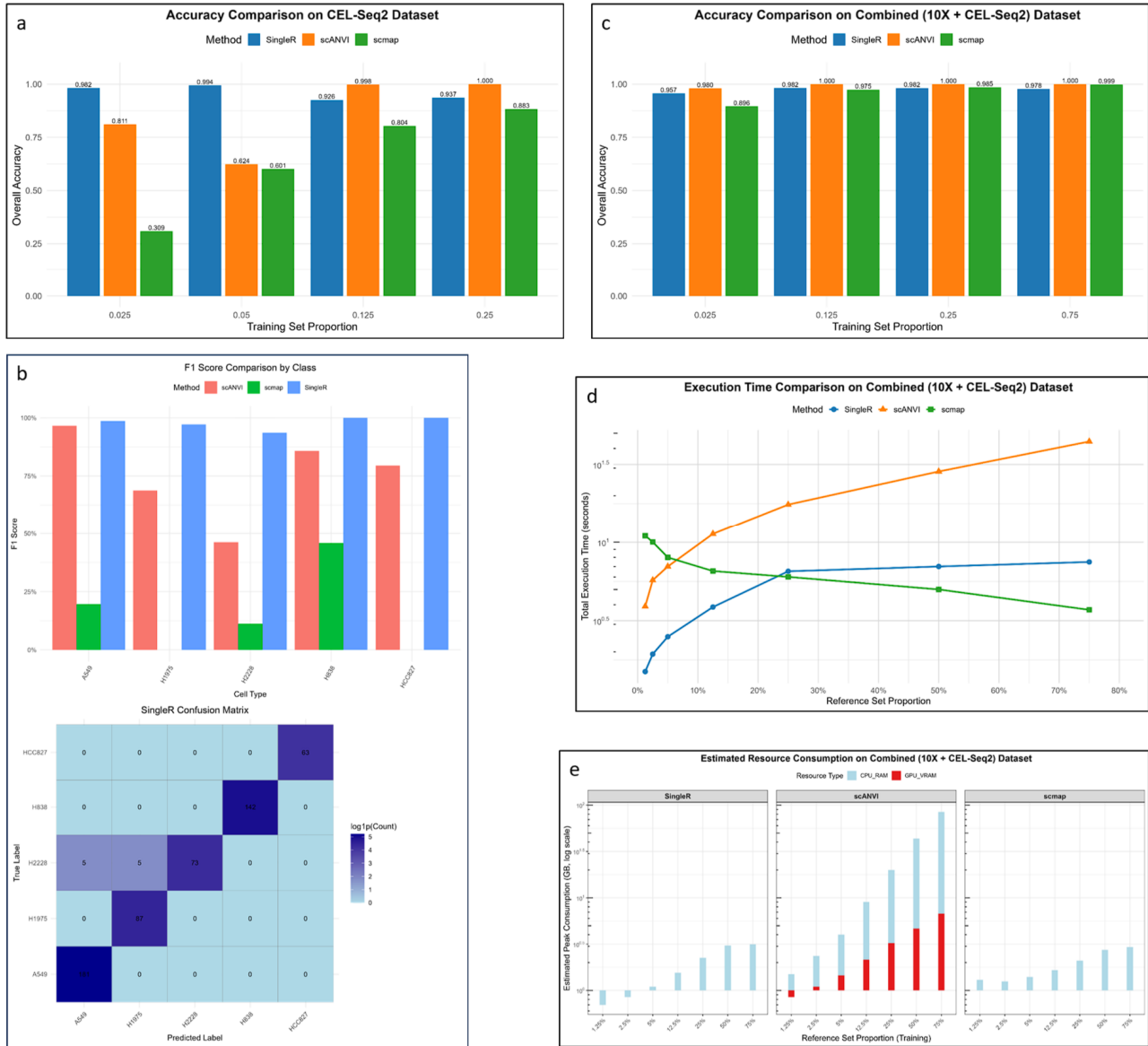


Figure 1 | Benchmarking automated cell type annotation tools across varying reference scales and dataset compositions.

a, Overall accuracy comparison on the CEL-Seq2 dataset (570 cells, 5 cell types) with decreasing reference set proportions (intra-dataset partitioning, from 25% down to 2.5%). **b**, Per-class F1-scores for the CEL-Seq2 dataset annotation using a 2.5% reference proportion, alongside the corresponding confusion matrix for SingleR under these conditions. **c**, Overall accuracy comparison on the combined 10x Chromium and CEL-Seq2 dataset (4373 cells, 5 cell types) with decreasing reference set proportions (from 75% down to 2.5%). **d**, Total execution time for SingleR, scmap-cell, and scANVI on the combined dataset across varying reference set proportions. **e**, Peak computational resource consumption on the combined dataset across varying reference set proportions. Stacked bars for scANVI show CPU RAM (blue) and GPU VRAM (red); bars for SingleR and scmap represent CPU RAM.

References

- Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., Butte, A. J., & Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2), 163–172.
- Kiselev, V. Y., Yiu, A., & Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. *Nature methods*, 15(5), 359–362.
- Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M. I., & Yosef, N. (2021). Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular systems biology*, 17(1), e9620.
- Cheng, C., Chen, W., Jin, H., & Chen, X. (2023). A Review of Single-Cell RNA-Seq Annotation, Integration, and Cell-Cell Communication. *Cells*, 12(15), 1970.
- Pasquini, G., Rojo Arias, J. E., Schäfer, P., & Busskamp, V. (2021). Automated methods for cell type annotation on scRNA-seq data. *Computational and structural biotechnology journal*, 19, 961–969.