CITER: A Cluster-Focused approach for Hierarchy Inference and Annotation of cell types applied to Latin-American Multimodal single-cell data

Felipe Gajardo [1], Danilo Ceschin [2,3], Marina Fernandez [2,3], Anna Lorenc [4], Julieth Lopez-Castiblanco [5], Diego Ramírez-Espinosa [6], Ana Laura Hernandez-Ledesma[6], Alejandra Schäfer[6], Tarran Rupall[4], Thais de Oliveira[4], Matiss Ozols[4], Carolina Alvarez[1], Liliana Lopez Kleine[7], Benilton de Sá Carvalho[8], Luis Tataje-Lavanda[9], Yesid Cuesta-Astroz [5], Pablo Romagnoli [2,3], Carla Jones [4], Gosia Trynka [4], Alejandra Medina-Rivera [6] & Marcela Sjoberg [1].

1. Laboratorio de Regulación Epigenética, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile (PUC), Chile.
2. Instituto Universitario de Ciencias Biomédicas de Córdoba (IUCBC).
3. Centro de Investigación en Medicina Traslacional "Severo R. Amuchástegui" (CIMETSA), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
4: Wellcome Sanger Institute, UK.
5: Instituto Colombiano de Medicina Tropical, Colombia.
6: Laboratorio Internacional de Investigación sobre el Genoma Humano (LIIGH-UNAM), Mexico.
7: Universidad Nacional de Colombia en Bogotá, Colombia.
8: Universidade Estadual de Campinas, Brazil.
9: Facultad de Ciencias de la Salud - EPMH. Universidad Privada San Juan Bautista (UPSJB), Peru.

Introduction

Single-cell research has made significant advances, driven by substantial improvements in flow cytometry and single-cell sequencing technologies. In particular, single-cell RNA sequencing (scRNA-seq) has emerged as a versatile tool for characterizing cellular heterogeneity and states at an unprecedented resolution. Moreover, multimodal approaches, such as CITE-seq, enable the simultaneous assessment of RNA and surface proteins on individual cells, allowing for a richer and more nuanced characterization of cell phenotypes.

The potential of these techniques is undeniable; however, the vast data they generate challenge traditional definitions of cell types, revealing a degree of heterogeneity higher than previously anticipated. Therefore, developing and rigorously evaluating multimodal analysis methods that integrate complex data types for cell type annotation is essential, as this will be crucial for constructing detailed cell atlases. These advances support the goals of the JAGUAR project to create an immune cell atlas of Latin American populations, where comprehensive characterization of cellular diversity is critical.

Here, we present and benchmark CITER, a novel reference-guided cell-type imputation method that leverages RNA and antibody-detected tags (ADT) from CITE-seq experiments. This method assumes that cell clusters inferred from multimodal scRNA-seq data accurately reflect cell heterogeneity, providing a basis for building cell-type hierarchies by selecting lineage-defining features at various levels of granularity. Additionally, CITER performs a marker-independent and cluster-focused quantile normalization to counteract unwanted biases associated with antibody affinities and cellular heterogeneity. Ultimately, we demonstrate that CITER effectively competes with state-of-the-art methods by inferring hierarchical structures and classifying cell types within peripheral blood mononuclear cell (PBMC) datasets, underscoring its potential to advance the JAGUAR project's goal of creating an immune cell atlas tailored to the unique diversity of Latin American populations.

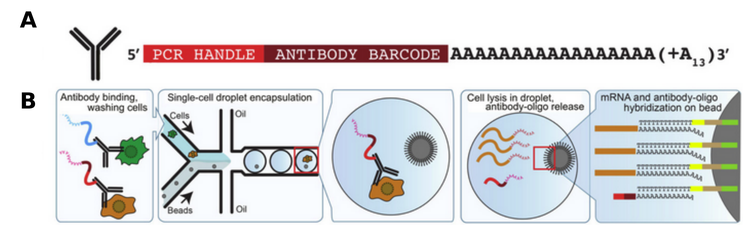


Figure 1. The CITEseq technology. (A) Representation of an antibody-conjugated oligonucleotide. (B) Schematic illustration of the key steps for capturing RNA (orange) and ADT (red) in CITE-seq libraries. Adapted from Choi et al., 2020.

Methodology

Table 1. Datasets used for training and testing purposes.

Dataset	# Cells	Subset size	# RNA features	# ADT features
PBMC reference (Hao et al., 2021)	161,764	34,153	20,729	228
COMBAT (Ahern et al., 2022)	836,148	835	32,063	192
JAGUAR pilot (this work)	3,551	-	53,335	137

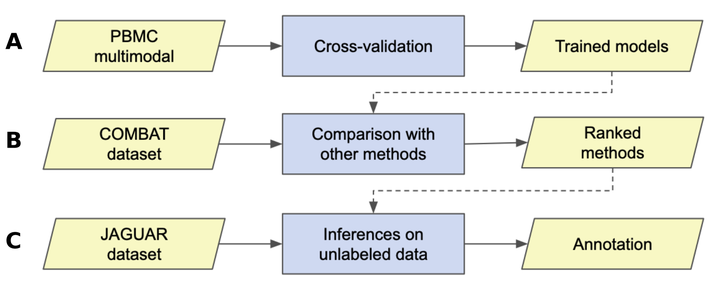


Figure 2. Strategy developed for the annotation of PBMCs for the JAGUAR project atlas. The flow-diagram depicts the CITE-seq based strategy for annotating cell types on JAGUAR generated data. (A) First, a cross-validation strategy is performed using a representative sample of the PBMC multimodal reference to evaluate the performance of CITER in ideal conditions. (B) CITER and other annotation methods (Azimuth, Superscan, and MMoCHI) are benchmarked using an independent reference dataset from the one used for training. (C) CITER and the other methods are used to annotate a real dataset from the JAGUAR project that has no curated annotation.

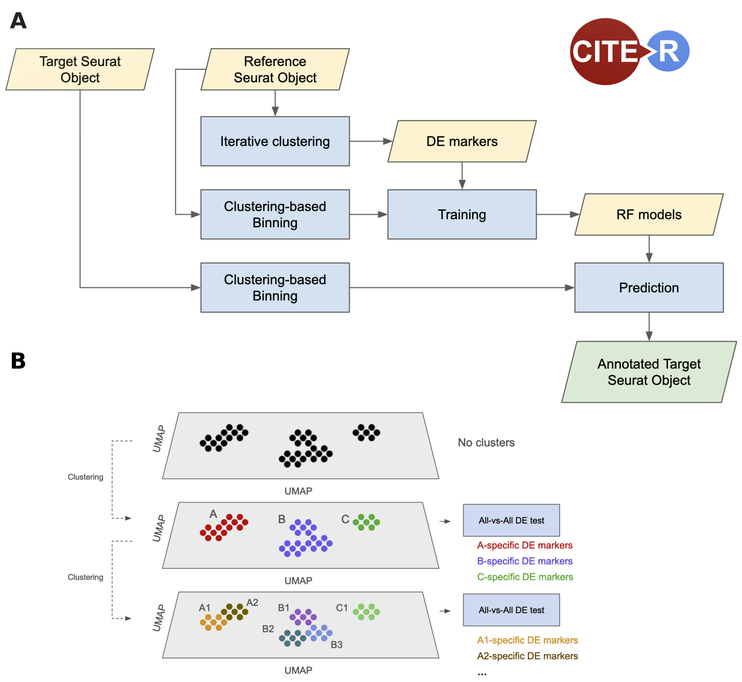


Figure 3. The CITER strategy for annotating cell types. (A) The flow-diagram depicts the CITER workflow. Briefly, informative RNA and ADT markers are infered from the reference dataset using our iterative clustering strategy (See B). Next, the matrices from the target and reference are quantile-normalized, and the latter is used to generate random forest models that ultimatelty will be used to predict cell-types into the target dataset. (B) Scheme illustrating the CITER interative clustering strategy.

Results

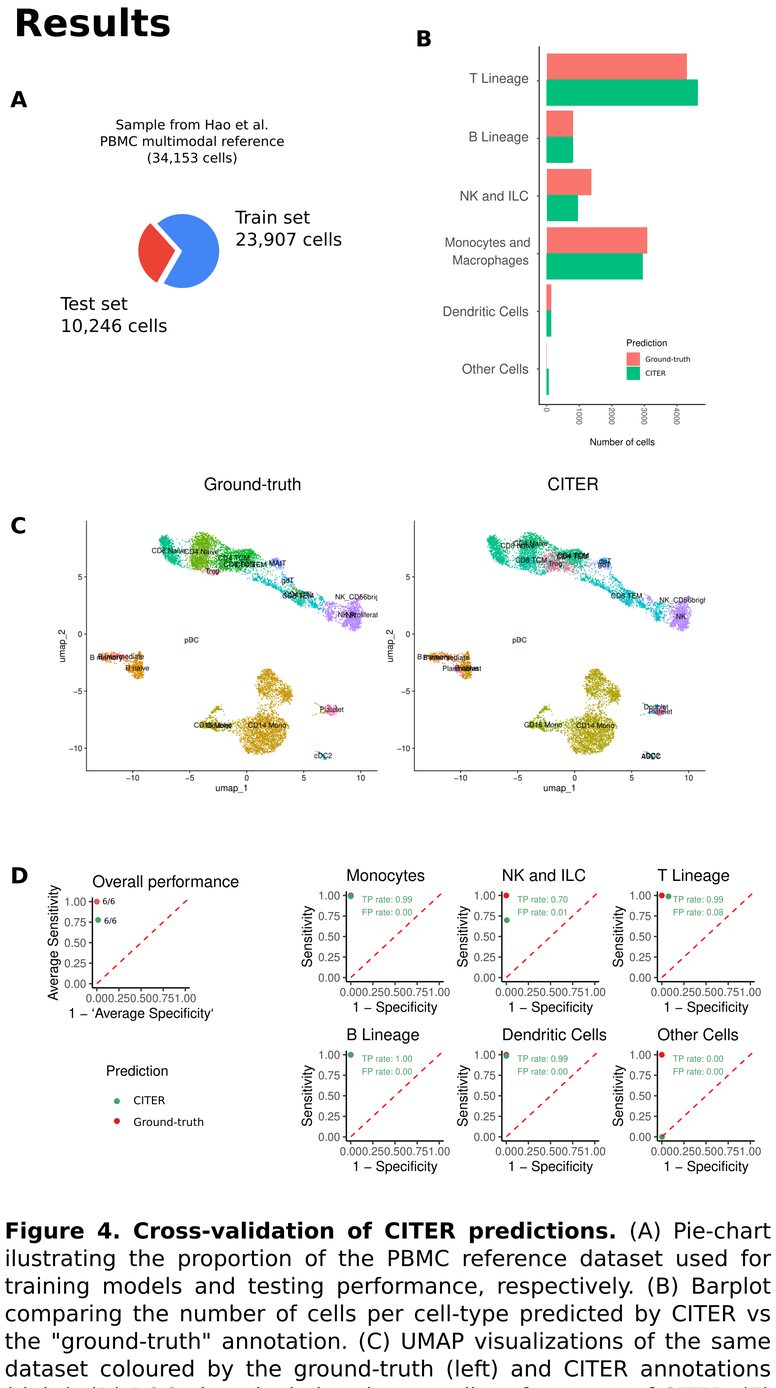
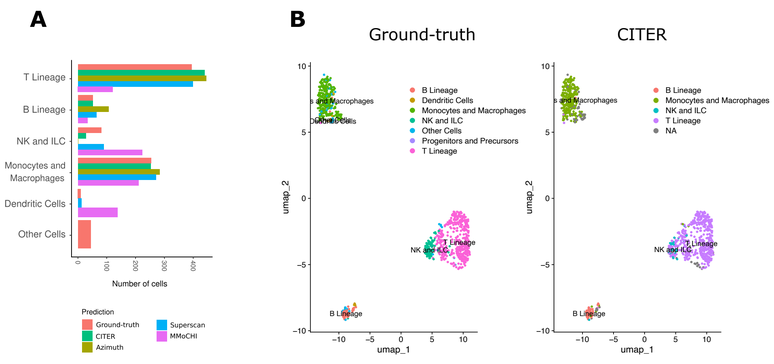


Figure 4. Cross-validation of CITER predictions. (A) Pie-chart ilustrating the proportion of the PBMC reference dataset used for training models and testing performance, respectively. (B) Barplot comparing the number of cells per cell-type predicted by CITER vs the "ground-truth" annotation. (C) UMAP visualizations of the same dataset coloured by the ground-truth (left) and CITER annotations (right). (D) ROC plots depicting the overall performance of CITER. (E) ROC plots depicting the cell-type specific performance of CITER.



Continues in the next panel...

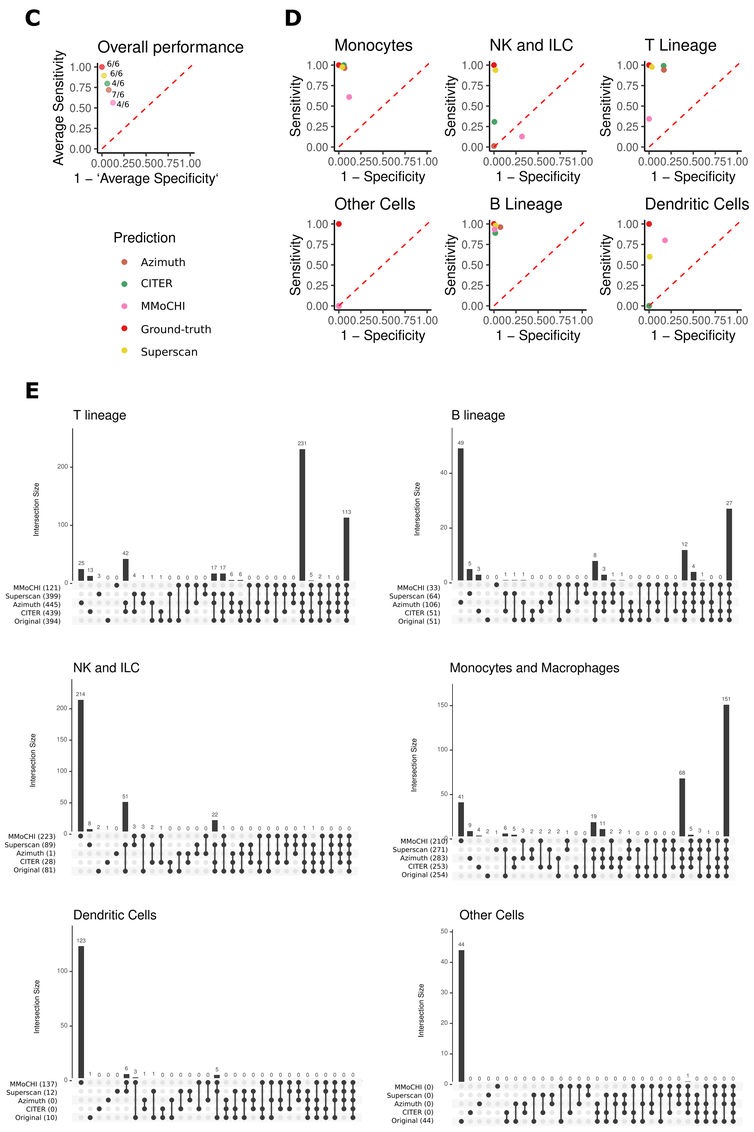


Figure 5. Comparison of multiple methods re-annotating the COMBAT dataset. (A) Barplot comparing the number of cells per cell-type predicted by the four methods evaluated. (B) UMAP plots contrasting the ground-truth and CITER annotations. (C) ROC plots depicting the overall performance of four different methods for annotating cell-types on single-cell sequencing data. (D) ROC plots depicting the cell-type specific performance of the four methods evaluated. (E) Upset plots detailing the number of cells with concordant annotations between the methods evaluated and the ground-truth.

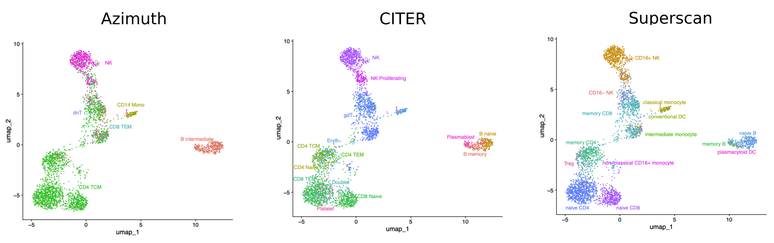


Figure 6. Annotation of multimodal data from Latin-American samples obtained by the JAGUAR project. Preliminary UMAP plots depicting the same data annotated using three different methods (Azimuth, CITER, and Superscan).

Conclusions

CITER can accurately predict cell-types using single-cell multimodal data. It competes and even outperforms other cell-type annotation methods.

There are cell-types that are more challenging to predict and show less concordance between annotation methods. In general, the method with the best performance across all cell types evaluated was Superscan, followed by CITER.

A combined strategy for cell-type annotation can certanly benefit single-cell atlas projects such as the JAGUAR.

Acknowledgements

This work is supported by the Chan Zuckerberg Initiative (CZI) for the Ancestry Networks of the Human Cell Atlas.

References

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., ... & Satija, R. (2021). Integrated analysis of multimodal single-cell data. Cell, 184(13), 3573-3587.

Ahern, D. J., Ai, Z., Ainsworth, M., Allan, C., Allcock, A., Angus, B., ... & Salio, M. (2022). A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. Cell, 185(5), 916-938.

Choi, J. R., Yong, K. W., Choi, J. Y., & Cowie, A. C. (2020). Single-cell RNA sequencing and its combination with protein and DNA analyses. Cells, 9(5), 1130.