Confident multimodal analysis of single cells across platforms and species

Alignment of single-cell proteomics data across platforms is difficult when different data sets contain limited shared features, as is typical in single-cell assays with antibody readouts. Therefore, we developed matching with partial overlap (MARIO) to enable confident and accurate matching for multimodal data integration and cross-species analysis.

This is a summary of:
Zhu, B. et al. Robust single-cell matching and multimodal analysis using shared and distinct features. *Nat. Methods* https://doi.org/10.1038/s41592-022-01709-7 (2023).

Published online:
Published online: 16 January 2023

The problem

Alignment of single-cell proteomics data from different sources is difficult when each experiment typically measures only a limited range of features, and particularly so when very few of these features overlap between different experiments. Many data integration tools are designed for high-parametric single-cell measurements, a characteristic that renders them unsuitable for addressing this problem[1]. We accordingly developed matching with partial overlap (MARIO), a cell matching algorithm that accounts for both shared and distinct features in data derived from different sources. MARIO is based on a novel one-way matching algorithm[2] and includes quality control steps to avoid suboptimal and biologically irrelevant matching. MARIO enables integrative, cross-modality and cross-species (such as non-human primate to human) analysis of single-cell protein data that could reveal new biological features and generate testable hypotheses.

The solution

MARIO leverages both the shared and distinct features of different data sets to optimize a global objective. Our algorithm uses linear assignment matching instead of a nearest neighbor approach to find local similarities. We also developed the matchability test and joint regularized filtering steps to prevent uninterpretable over-integration. The matchability test reveals the strength of underlying information connecting the datasets whereas the joint regularized filtering step removes suboptimal cell pairing even in the absence of prior knowledge. Benchmarking of MARIO on multiple single-cell proteomic platforms – cytometry by time of flight (CyTOF), cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) and co-detection by indexing (CODEX) – revealed a consistent first-in-class performance over existing methods (Fig. 1a).

We next applied MARIO to integrate data from CODEX and CITE-seq experiments, which uncovered new and spatially orchestrated immunological interactions between complement-expressing macrophages and neutrophils from patients with COVID-19. These results were validated using our multimodal spatial methodology[3] and shown to be highly concordant (Fig. 1b). We expect that the MARIO methodology will be a useful resource for the cell biology community when applied to various experimental platforms and biological processes.

The implications

The ability to infer unmeasured features at the single-cell level has been a long-standing goal of single-cell research. MARIO now enables confident cross-modality inference of unmeasured features at the single-cell level, such as transcriptional profiles measured with the spatial proteomics platform CODEX.

A key limitation of MARIO is the need for partially overlapping features to enable confident cell matching. Absence of overlap in the available markers can result in heavily biased results. For example, the lack of B cell-specific markers in one of the datasets is likely to impair the inference of B cell-related biological processes beyond those included as measured features in other data sets. As such, implementation of the two quality control steps is essential to reduce user over-interpretation of MARIO-derived results.

We foresee an important area of focus in the future being the matching and inference of cell features lying beyond the immediately overlapping markers that have actually been measured. For example, MARIO could aid in attempts to infer protein expression levels from RNA features.

Bokai Zhu[1] & Sizun Jiang[2]
[1]Stanford University, Palo Alto, CA, USA.
[2]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA.

EXPERT OPINION

"Given the growing number of data sets of diverse modalities, algorithms that enable the integration of such data are of high value. The authors developed MARIO to perform this difficult task. They applied it to integrate CyTOF, CITE-seq and CODEX datasets in several scenarios spanning healthy unperturbed cells, stimulated cells and specimens from patients with COVID-19, as well as across species, and show several metrics comparing MARIO to previously published methods." **Karin Pelka, Gladstone Institutes, San Francisco, CA, USA.**

FIGURE

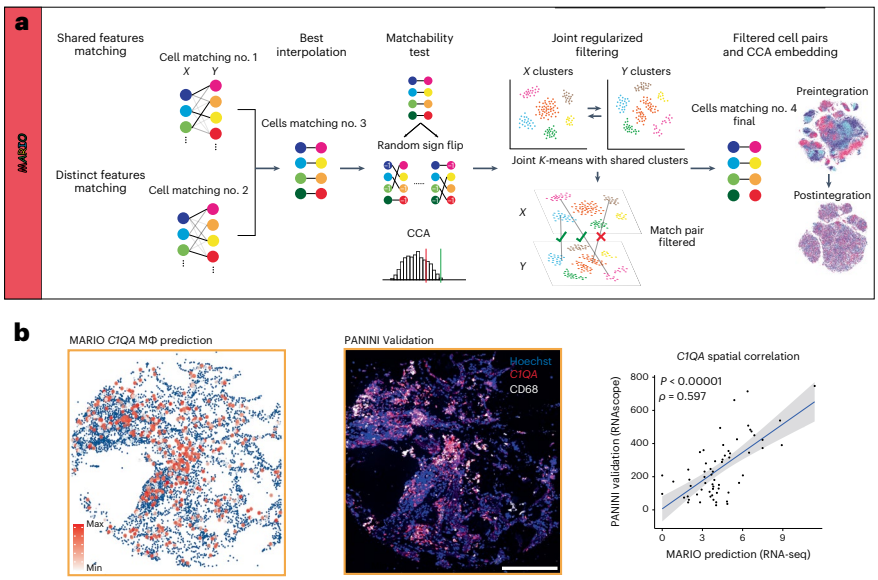

Fig. 1 | MARIO method and experimental validation. a, MARIO pipeline overview, including initial matching, refined matching, interpolation, matchability test and joint regularized filtering. **b**, MARIO-inferred CITE-seq *C1Q* RNA expression (left), experimentally validated *C1Q* expression (center) on non-serial sections, and dot plot (right) of spatial correlations between MARIO-predicted and experimental data (lines, 95% CI; *P* and *ρ* values calculated by two-sided Spearman ranked test). Scale bar, 400 µm. CCA, canonical correlation analysis.

BEHIND THE PAPER

The two lead authors of this paper, Bokai Zhu and Shuxiao Chen, have been close friends since college, where they took statistics and biology classes together. Among their many initially naive discussions was the idea of comparing organs and tissues across species with a statistically rigorous method. After graduating from Cornell University, they parted ways to undertake their PhD training on different US coasts. Fortuitously, they were reunited several years later and applied their complementary expertise to address some of those early ideas. This work led to the development of MARIO, a step towards realization of their initial dream. Bokai and Shuxiao are grateful for this enjoyable collaborative experience and sincerely hope that the tool they created can be helpful to the cell biology field. **B.Z.**

REFERENCES

1. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
   This comprehensive benchmarking paper describes different integration methods.
2. Chen, S. et al. One-way matching of datasets with low rank signals. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2204.13858.
   This preprint paper reports one of the key algorithms that powers the MARIO method.
3. Jiang, S. et al. Combined protein and nucleic acid imaging reveals virus-dependent B cell and macrophage immunosuppression of tissue microenvironments. *Immunity* **55**, 1118–1134.e8 (2022).
   This paper reports the co-staining technique PANINI.

FROM THE EDITOR

"MARIO is a method to integrate single-cell proteomics datasets by leveraging both shared and unique features of cell types. We expect that MARIO will serve as a useful and enabling methodology to assess complex protein expression data at a high level." **Madhura Mukhopadhyay, Senior Editor, *Nature Methods*.**