Bioinformatics

https://doi.org/10.1038/s43588-023-00489-0

Cellular deconvolution with continuous transitions

Zheyang Zhang & Jialiang Huang

A recent work introduces a cellular deconvolution method, MeDuSA, of estimating cell-state abundance along a one-dimensional trajectory from bulk RNA-seq data with fine resolution and high accuracy, enabling the characterization of cell-state transition in various biological processes.

Tissues are complex ecosystems composed of various morphologically different and functionally specialized cell types. Furthermore, cells of the same type can exhibit multiple states in different anatomical regions, disease conditions and genetically distinct individuals, making the cell composition highly heterogeneous. Quantifying the proportion of cell types (or states) within a tissue of interest is critical to gaining insights into basic biology and medicine (for example, human development and cancer)[1]. Attempts to elucidate cell composition at a fine-grained level using traditional methods such as flow cytometry and immunohistochemistry often fail, owing to the fact that these methods reply on fewer preselected markers that only cover known coarse-grained cell types. Although single-cell RNA sequencing (scRNA-seq) enables unbiased transcriptional profiling of thousands of individual cells, it is not cost-effective for large-scale cohorts. For these reasons, many cellular deconvolution methods have been developed to infer cell-type abundance from bulk RNA-seq or microarray data over the past two decades[2,3], thus providing a cost- and

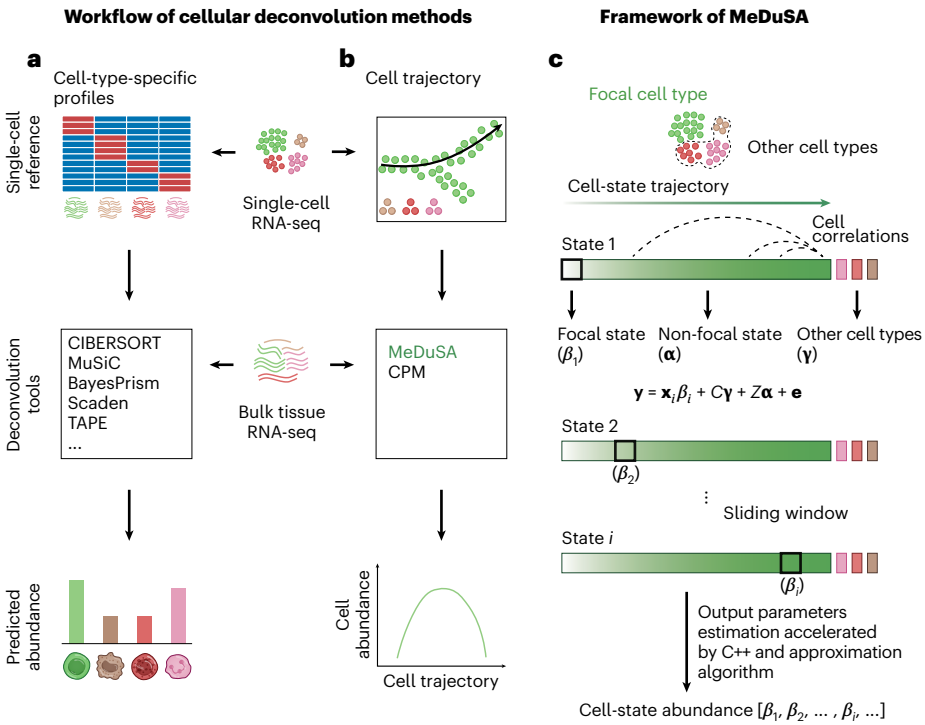

Fig. 1 | The overview of various cellular deconvolution methods. a, Workflow for cellular deconvolution methods that use cell-type-specific profiles derived from scRNA-seq data as reference, and output predicted cell-type abundance. **b**, Workflow for cellular deconvolution methods that exploit the continuous dynamic transitions of single cells, and output fine-resolution cell-state abundance along a one-dimensional trajectory. **c**, The framework of MeDuSA. The linear mixed model (LMM) is characterized by the formula $\mathbf{y} = \mathbf{x}_i\beta_i + C\gamma + Z\alpha + \mathbf{e}$, where $\mathbf{y}$ is a vector comprising the expression levels of signature genes (associated with cell-state trajectory) in the bulk RNA-seq data. Given a cell state i, $\mathbf{x}_i$ represents the expression levels of signature genes in cells in the focal state in scRNA-seq data, while β_i represents the cell-state abundance to be estimated. C is the gene expression matrix, with each row representing a signature gene and each column representing the mean of each of the other cell types, while $\boldsymbol{\gamma}$ is a vector of the corresponding effects. Z is also a gene expression matrix, with each row representing a signature gene and each column representing each of the cells at non-focal state, while $\boldsymbol{\alpha}$ is a vector of the corresponding effects. $\mathbf{e}$ is the residual. In this model, β_i and $\boldsymbol{\gamma}$ are treated as fixed effects, whereas $\boldsymbol{\alpha}$ is treated as a random effect.

time-effective way to bridge cell composition and clinical significance (Fig. 1a). An Article in this issue of *Nature Computational Science* introduces a cellular deconvolution method — mixed model-based deconvolution of cell-state abundance (MeDuSA) — that estimates cell-state abundance along a one-dimensional trajectory[4]. In this paper, Song and colleagues use a linear mixed model (LMM) that combines fixed- and random-effect terms, greatly improving deconvolution accuracy over existing methods.

Single-cell transcriptomic techniques continue to revolutionize the resolution of cell analysis, determining discrete cell types and cell states with continuous dynamic transitions that can be related to development and disease progression[5]. Cells in different states can be computationally ordered according to a pseudo-time series, or cell trajectory[6]. Both MeDuSA and another method, Cell Population Mapping (CPM)[7], were developed to exploit the rich spectrum of single-cell reference profiles to estimate cell-state abundance in bulk RNA-seq data, which enables fine-resolution cellular deconvolution (Fig. 1b). Although CPM effectively tackles the issue of estimating the abundance of cells in different states, MeDuSA further improves the estimation accuracy by employing a LMM (see the equation in Fig. 1c) that takes into account both the cell state of interest (focal state) and the remaining cells of the same cell type (non-focal state) as well as the other cell types. More specifically, MeDuSA iteratively incorporates focal-state cells along a cell-state trajectory as a fixed effect, while treating non-focal-state cells individually as random effects to account for correlations among them (Fig. 1c). The random-effect component allows the incorporation of individualized weights for each cell. This is beneficial because it acknowledges that different cells may contribute differently to the overall gene expression observed in bulk RNA-seq data. By assigning specific weights to each cell, the model can more accurately capture the variability in gene expression. Meanwhile, MeDuSA benefits from a LMM that mitigates the collinearity problem[8], which can pose obstacles in linear regression tasks, between focal state and non-focal state cells. Furthermore, the core algorithm is coded in C++ and includes an approximation for estimating LMM parameters, substantially improving its computational efficacy.

The authors demonstrate the diverse application scenarios of MeDuSA by showcasing the correlation between cell-state abundance and different biological conditions: disease conditions (such as esophageal carcinoma versus normal esophagus), clinical outcomes (such as clinical phases in COVID-19 and survival time in skin melanoma), mechanisms of pathogenicity (such as the TCR expansion level in melanoma), and treatment exposures (such as response to anti-PD-1 in skin melanoma). They also identify cell-state-dependent genetic control of transcription by applying MeDuSA to the GTEx tissues dataset, a cohort that has both genotype and bulk RNA-seq information. We note that this has only recently been achieved with cohort-level scRNA-seq data. Collectively, these cases exemplify how cellular deconvolution, particularly at the fine resolution and high accuracy achieved by MeDuSA, has accelerated our comprehension in the realms of basic biology and medicine.

Despite its usefulness, MeDuSA has certain limitations that should be taken into consideration. First, MeDuSA relies on a pre-annotated cell-state trajectory, and different trajectory inference methods can introduce different trajectories. This can introduce inconsistent estimates of cell-state abundance. Second, MeDuSA considers only a one-dimensional trajectory, and thus, for deconvolving increasingly intricate cell-state (two- or multi-dimensional trajectories), MeDuSA will need to be improved. Third, some cell types (such as dendritic cells and regulatory T cells) are relatively rare but important in controlling the human immune response, and MeDuSA's ability to estimate such rare cell-state abundances needs to be established. Last, cellular deconvolution methods including MeDuSA focus solely on known cell types and states, disregarding the existence of unknowns. This gives rise to the more challenging task of regressing out the effects of unknown cell types or states. Once these limitations have been addressed, such cellular deconvolution methods should pave the way for more accurate and comprehensive analyses of biological systems, leading to deeper insights into cellular heterogeneity and functional dynamics.

Zheyang Zhang[1,2] & Jialiang Huang[1,2] ✉
[1]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Faculty of Medicine and Life Sciences, Xiamen University, Xiamen, Fujian, China. [2]National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, Fujian, China.
✉e-mail: jhuang@xmu.edu.cn

References
1. Newman, A. M. et al. *Nat. Biotechnol.* **37**, 773–782 (2019).
2. Jin, H. & Liu, Z. *Genome Biol.* **22**, 102 (2021).
3. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. *Nat. Commun.* **11**, 5650 (2020).
4. Song, L., Sun, X., Qi, T. & Yang, J. *Nat. Comput. Sci.* https://doi.org/10.1038/s43588-023-00487-2 (2023).
5. Wagner, A., Regev, A. & Yosef, N. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
6. Trapnell, C. et al. *Nat. Biotechnol.* **32**, 381–386 (2014).
7. Frishberg, A. et al. *Nat. Methods* **16**, 327–332 (2019).
8. Allen, M. P. *Understanding Regression Analysis* 176–180 (Springer, 1997).

Competing interests
The authors declare no competing interests.