University of Pittsburgh
School of Public Health

Accurate estimation of rare cell type fractions from tissue omics data via hierarchical deconvolution

Penghui Huang[1], Manqi Cai[1], Xinghua Lu[2], Chris McKennan[3] and Jiebiao Wang[1]
[1]Department of Biostatistics, [2]Department of Biomedical Informatics, [3]Department of Statistics, University of Pittsburgh

Introduction

- The varying cellular fractions of tissue samples serve as the foundation for many downstream statistical analyses. Although biochemical methods can measure cell counts of samples, they are labor-intensive and costly.
- Cellular deconvolution methods have been developed to estimate cellular proportions. Supervised methods can be modeled as

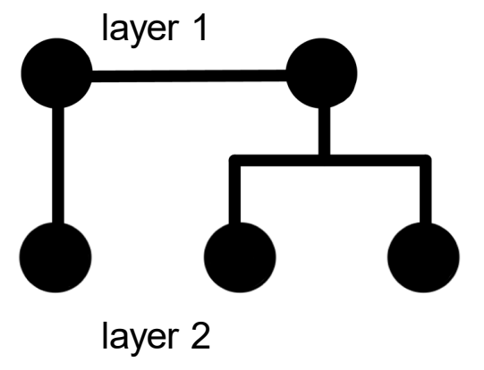
$$\underset{m\times n}{\overset{\text{Bulk}}{\boldsymbol{X}}} = \underset{m\times K}{\overset{\text{Signature}}{\boldsymbol{A}}}\,\underset{K\times K}{\overset{\text{Cell size}}{\boldsymbol{S}}}\,\underset{K\times n}{\overset{\text{CT fractions}}{\boldsymbol{P}}} + \underset{m\times n}{\boldsymbol{E}}$$

- **m** marker genes
- **n** tissue samples
- **K** cell types

- Cell type hierarchy has become important for understanding the topology of cell types across datasets[1,2,3].
- The increase of number of cell types begets co-linearity because of shared origins of cell differentiation and leads to rare cell types.
- Here we present Hierarchical Deconvolution (HiDecon), a penalized model pooling information across related cell types to tackle these challenges.

Estimation Model

- Given hierarchical tree, CT fractions of "parent" and "children" across layers have a summation relationship approximately.
- We introduce a cell type mapping matrix $\boldsymbol{B}_{l,(l+1)} \in \mathbb{R}_+^{K_l \times K_{l+1}}$ modeling the relationship between layer l and $l+1$. e.g.,



$$\boldsymbol{p}_{i1} \approx \boldsymbol{B}_{1,2}\boldsymbol{p}_{i2} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix} \boldsymbol{p}_{i2}$$

- HiDecon model objective function (sample i) for an L layer tree:

$$f(\boldsymbol{p}_i) = \sum_{l=1}^{L} m_l^{-1} \|\boldsymbol{x}_{il} - \boldsymbol{A}_l \boldsymbol{S}_l \boldsymbol{p}_{il}\|_2^2 + \lambda \sum_{l=1}^{L-1} K_l^{-1} \|\boldsymbol{p}_{il} - \boldsymbol{B}_{l,(l+1)} \boldsymbol{p}_{i(l+1)}\|_2^2$$
$$\boldsymbol{p}_{il} \geq 0, \quad \|\boldsymbol{p}_{il}\|_1 = 1.$$

where p_i is a length $K = \Sigma_{l=1}^{L} K_l$ vector of cell proportions of all the nodes in the hierarchical tree, p_{il} denotes the layer l part of type fractions, x_{il} denotes bulk gene expression level of m_l markers in layer l, A_l and S_l are reference signature matrix and size factor matrix derived from single cell data respectively. Marker genes are selected for each layer.

- To estimate fractions for all layer simultaneously, we rewrite the model as:

$$f(\boldsymbol{p}_i) = \|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{A}}\boldsymbol{p}_i\|_2^2 + \lambda \|\tilde{\boldsymbol{B}}\boldsymbol{p}_i\|_2^2, \quad \text{such that } \boldsymbol{p}_i \geq 0 \text{ and } \|\boldsymbol{p}_{il}\|_1 = 1$$

$$\tilde{\boldsymbol{x}} = \begin{pmatrix} m_1^{-1/2}\boldsymbol{x}_{i1} \\ \vdots \\ m_L^{-1/2}\boldsymbol{x}_{iL} \end{pmatrix}, \quad \tilde{\boldsymbol{A}} = \begin{pmatrix} m_1^{-1/2}\boldsymbol{A}_1\boldsymbol{S}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & m_L^{-1/2}\boldsymbol{A}_L\boldsymbol{S}_L \end{pmatrix} \quad \tilde{\boldsymbol{B}} = \begin{pmatrix} K_1^{-1/2}I_{K_1} & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & K_{L-1}^{-1/2}I_{K_{L-1}} & 0 \end{pmatrix} - \begin{pmatrix} 0 & K_1^{-1/2}B_{1,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & K_{L-1}^{-1/2}B_{(L-1),L} \end{pmatrix}$$

$$\boldsymbol{H}(\boldsymbol{p}_i) = \nabla^2_{\boldsymbol{p}_i} f(\boldsymbol{p}_i) = \tilde{\boldsymbol{A}}^T\tilde{\boldsymbol{A}} + \lambda\tilde{\boldsymbol{B}}^T\tilde{\boldsymbol{B}} \qquad \boldsymbol{b}_k = \tilde{\boldsymbol{A}}^T\tilde{\boldsymbol{x}}_i$$

HiDecon estimation algorithm:

(Coordinate-wise descend algorithm[4])

```
1. Initialize $\boldsymbol{p}_i = \boldsymbol{H}^{-1}\boldsymbol{b}$.
2. If $\boldsymbol{p}_i \geq 0$, return $\boldsymbol{p}_i$. If not, proceed to step 3.
3. Let $p_{ik} = max\{p_{ik}, 0\}$, for all $k \in \{1, \cdots, K\}$.
4. $p_{ik} = max\left\{0, p_{ik} + \frac{\boldsymbol{b}_k - \boldsymbol{p}_i^T\boldsymbol{H}_{*k}}{H_{kk}}\right\}$, where $\boldsymbol{H}_{*k}$ is the $k$th column of $\boldsymbol{H}$.
5. Check the KKT conditions
       $\left(|\boldsymbol{H}_{*k}^T\boldsymbol{p}_i - \boldsymbol{b}_k| \leq \epsilon\right)$ OR $\left(\boldsymbol{H}_{*k}^T\boldsymbol{p}_i - \boldsymbol{b}_k \geq 0 \text{ AND } p_{ik} \leq \epsilon\right), \forall k \in \{1, \cdots, K\}$.
6. If KKT are satisfied, return $\boldsymbol{p}_i$. Otherwise, repeat step 4-5.
```

Parameter Selection

- For tuning parameter λ, we employ a resampling method shown as below:

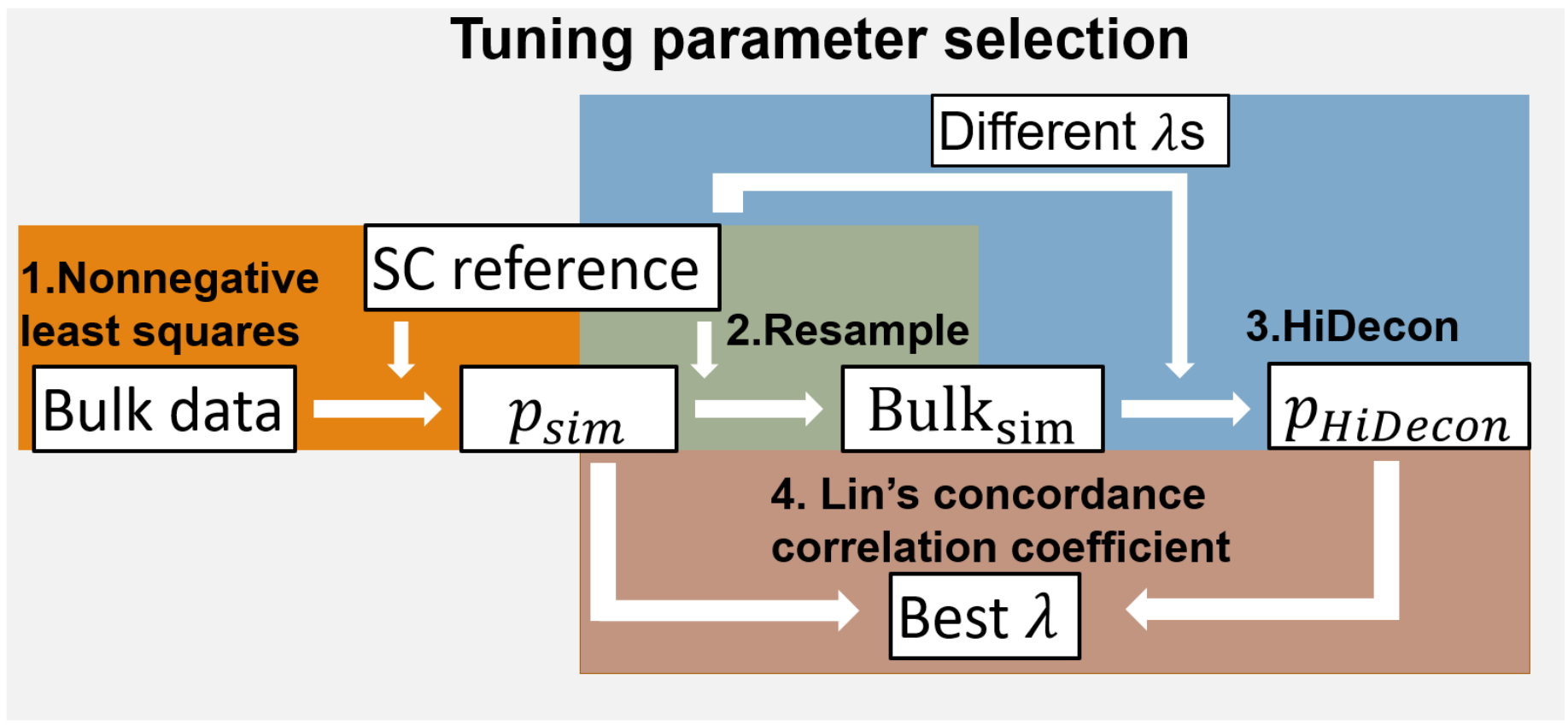


Figure 1: Flow chart for HiDecon tuning parameter selection.

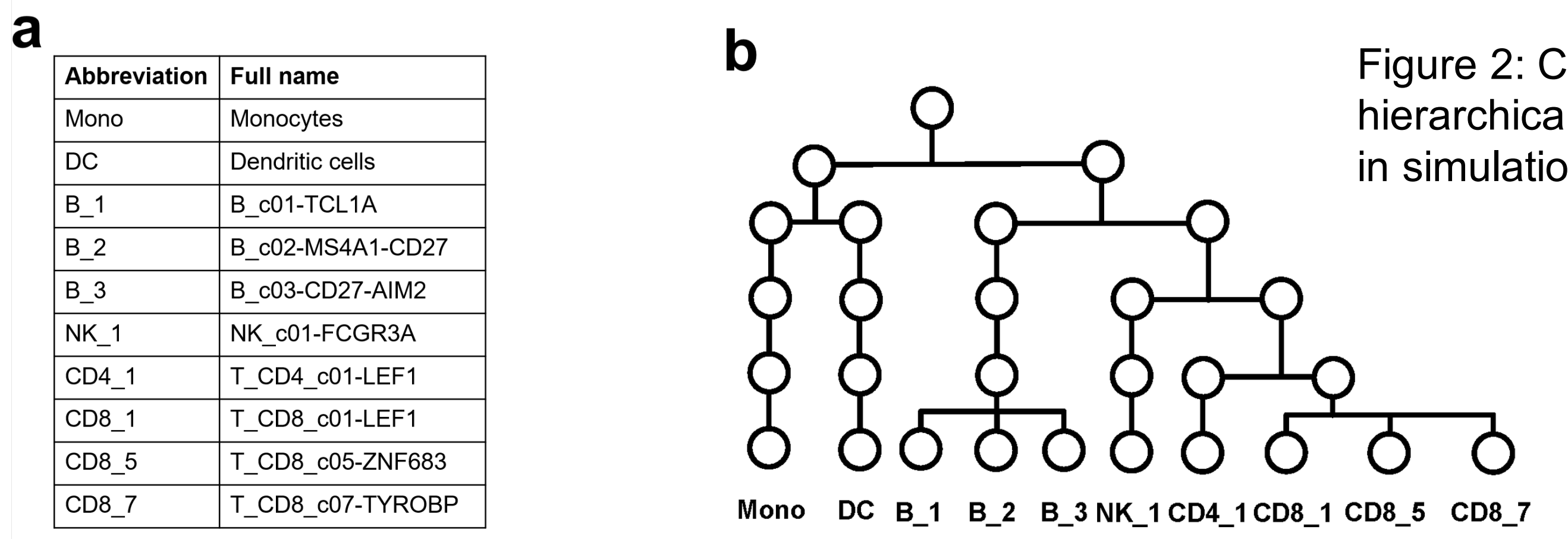
Simulation Studies

Metrices

- MAE: mean absolute error compared with true fraction $avg\left(|\boldsymbol{P} - \hat{\boldsymbol{P}}|\right)$
- CCC (Lin's Concordance Correlation Coefficient[5]): $\frac{2cov\left(\boldsymbol{P}_{k*}, \hat{\boldsymbol{P}}_{k*}\right)}{\sigma^2_{\boldsymbol{P}_{k*}} + \sigma^2_{\hat{\boldsymbol{P}}_{k*}} + \left(avg\left(\boldsymbol{P}_{k*} - \hat{\boldsymbol{P}}_{k*}\right)\right)^2}$

Data: real large scale COVID-19 individual level scRNA-seq dataset[6] of PBMC (peripheral blood mononuclear cells) with subtypes:

- Simulated bulk data: averaged across single cells within sample (126 samples).
- Reference: all individual level single cell data used for simulating bulk data.



Abbreviation	Full name
Mono	Monocytes
DC	Dendritic cells
B_1	B_c01-TCL1A
B_2	B_c02-MS4A1-CD27
B_3	B_c03-CD27-AIM2
NK_1	NK_c01-FCGR3A
CD4_1	T_CD4_c01-LEF1
CD8_1	T_CD8_c01-LEF1
CD8_5	T_CD8_c05-ZNF683
CD8_7	T_CD8_c07-TYROBP

Figure 2: Cell types and hierarchical tree we use in simulation.

	Mono	DC	B_1	B_2	B_3	NK_1	CD4_1	CD8_1	CD8_5	CD8_7	Mean	MAE
	(0.38)	(0.02)	(0.07)	(0.03)	(0.03)	(0.07)	(0.13)	(0.09)	(0.09)	(0.08)	CCC	
HiDecon	0.85	0.25	**0.80**	**0.54**	**0.50**	0.34	**0.62**	0.66	0.52	0.64	**0.57**	**0.05**
CIBERSORT	0.88	**0.31**	0.66	0.31	0.37	0.01	0.42	**0.76**	0.25	0.25	0.42	0.07
dtangle	0.35	0.08	0.56	0.18	0.30	**0.47**	0.51	0.51	**0.60**	**0.69**	0.43	0.06
MuSiC	**0.89**	0.01	NA	NA	0.16	NA	NA	0.35	0.22	0.39	0.20	0.08

Table 1: CCC and MAE in the simulation study for different deconvolution methods.

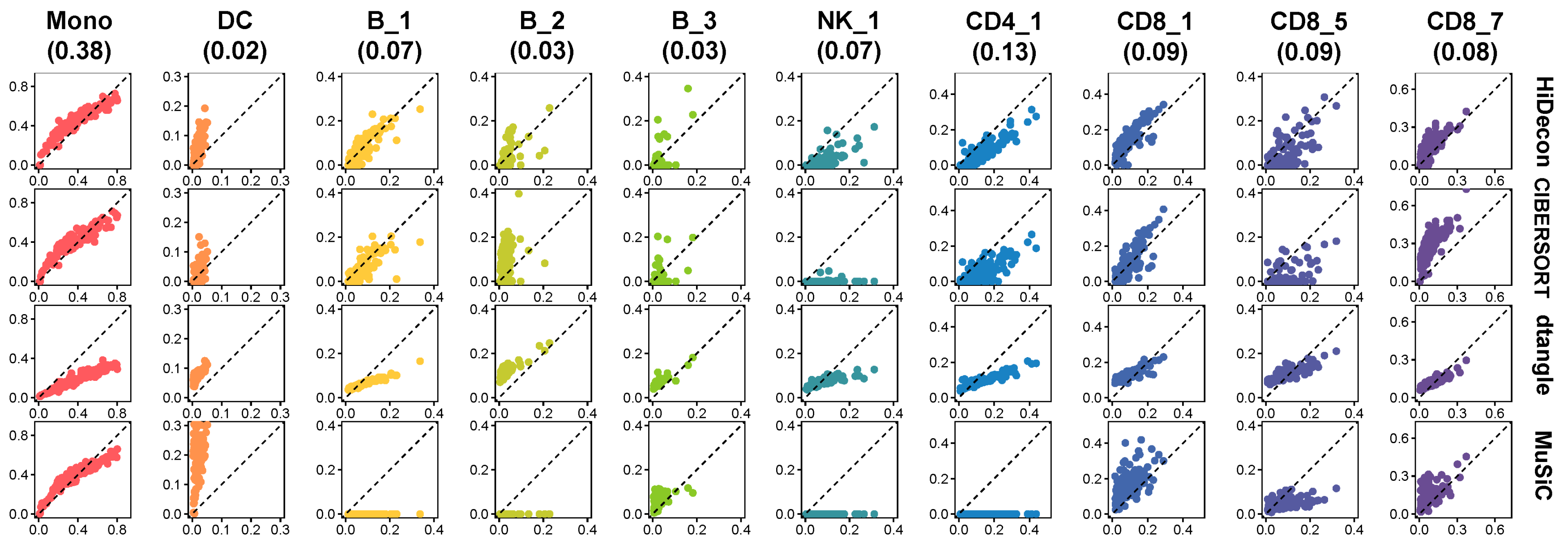


Figure 3: Scatter plots of cellular fractions in simulation study for different methods

Robustness analysis: We use simulated data and add noises $N(0, sd^2)$ to bulk data with maximum sd equals the sd of bulk data. Experiments are repeated 50 times using different random seeds and we averaged metrics among repetitions.

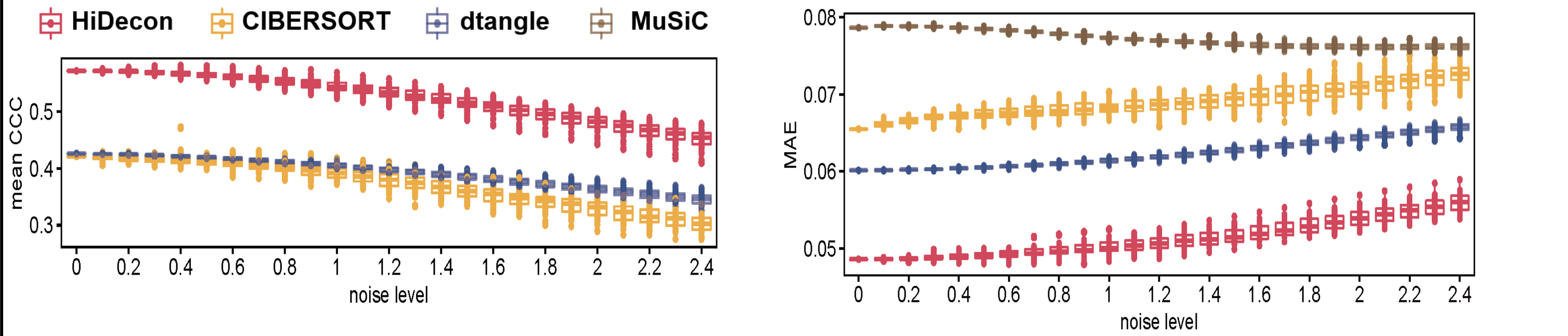


Figure 4: CCC (upper panel) and MAE in robustness analysis. (MuSiC CCC missing due to all 0 estimates).

Real Data Applications

Real data application: FHS data (Framingham Heart Study)[7,8,9] . Human blood data.

- Bulk data: 4,110 blood samples with measured cell type fractions.
- Reference: LM22 data (microarray data).
- Ground truth: measured blood cell counts in FHS.

	Neutrophil	Lymphocyte	Monocyte	Eosinophil	Mean	MAE
	(0.60)	(0.28)	(0.09)	(0.03)	CCC	
HiDecon	0.13	**0.57**	**0.04**	**0.28**	**0.26**	**0.10**
CIBERSORT	**0.15**	0.31	0.02	0.06	0.13	0.15
dtangle	0.02	0.17	0.01	0.01	0.05	0.17
HEpiDISH	0.12	0.25	0.03	0.04	0.11	0.17
MuSiC	NA	0.08	0.00	NA	0.02	0.32

Table 2: CCC and MAE in FHS data study for methods.

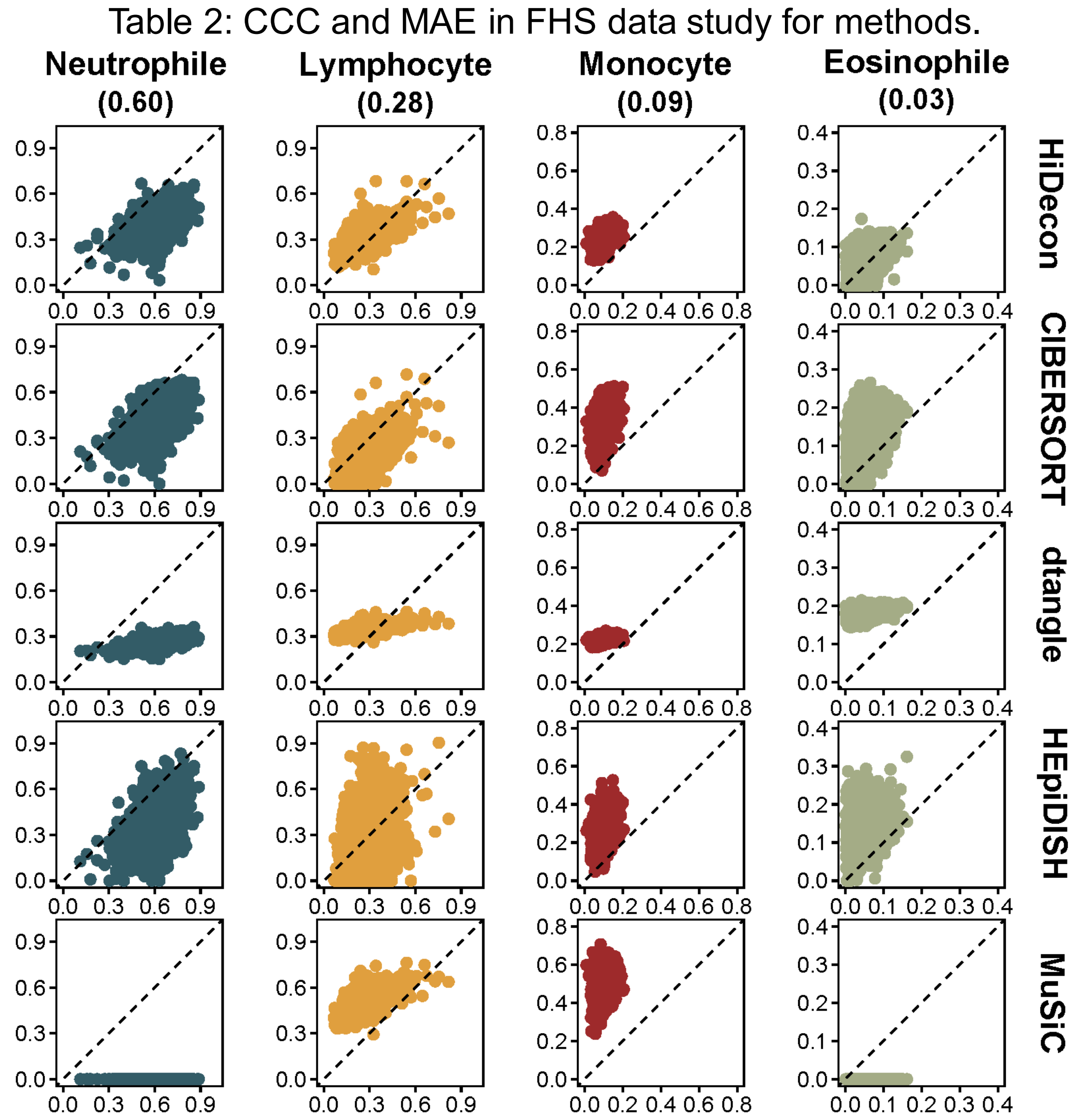


Figure 5: Scatter plots in the FHS real data study for different deconvolution methods.

Conclusions

- We developed HiDecon to incorporate a hierarchical cell type tree to facilitate the estimation of related cell types.
- HiDecon can incorporate complex tree structure with more flexibility.
- HiDecon can provide accurate estimates especially for rare cell types.
- We offer a user-friendly R package along with a brief tutorial hosted on https://github.com/randel/HiDecon.

References

1. Wu, Zhijin, and Hao Wu. "Accounting for cell type hierarchy in evaluating single cell RNA-seq clustering." *Genome biology* 21, no. 1 (2020): 1-14.
2. Peng, Minshi, Brie Wamsley, Andrew G. Elkins, Daniel H. Geschwind, Yuting Wei, and Kathryn Roeder. "Cell type hierarchy reconstruction via reconciliation of multi-resolution cluster tree." *Nucleic acids research* 49, no. 16 (2021): e91-e91.
3. Chen, Luxiao, Ziyi Li, and Hao Wu. "Cedar: incorporating cell type hierarchy improves cell type specific differential analyses in bulk omics data." *bioRxiv* (2022): 2022-07.
4. Franc, Vojtěch, Václav Hlaváč, and Mirko Navara. "Sequential coordinate-wise algorithm for the non-negative least squares problem." In *Computer Analysis of Images and Patterns: 11th International Conference, CAIP 2005, Versailles, France, September 5-8, 2005. Proceedings 11*, pp. 407-414. Springer Berlin Heidelberg, 2005.
5. Lawrence, I., and Kuei Lin. "A concordance correlation coefficient to evaluate reproducibility." *Biometrics* (1989): 255-268.
6. Ren, Xianwen, Wen Wen, Xiaoying Fan, Wenhong Hou, Bin Su, Pengfei Cai, Jiesheng Li et al. "COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas." *Cell* 184, no. 7 (2021): 1895-1913.
7. Dawber, Thomas R., Gilcin F. Meadors, and Felix E. Moore Jr. "Epidemiological approaches to heart disease: the Framingham Study." *American Journal of Public Health and the Nations Health* 41, no. 3 (1951): 279-286.
8. Feinleib, Manning, William B. Kannel, Robert J. Garrison, Patricia M. McNamara, and William P. Castelli. "The Framingham offspring study. Design and preliminary data." *Preventive medicine* 4, no. 4 (1975): 518-525.
9. Splansky, Greta Lee, Diane Corey, Qiong Yang, Larry D. Atwood, L. Adrienne Cupples, Emelia J. Benjamin, Ralph B. D'Agostino Sr et al. "The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination." *American journal of epidemiology* 165, no. 11 (2007): 1328-1335.

Acknowledgement

This research was funded in part through NIH's R01AG080590, R03OD034501, and R01MH123184, and a grant from the University of Pittsburgh UPMC Health System's Competitive Medical Research Fund. This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195, HHSN268201500001I and 75N92019D00031). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI.