SSMD: A semi-supervised approach for a robust cell type identification and deconvolution of mouse transcriptomics data

Xiaoyu Lu[1, 2+], Szu-Wei Tu[1, 2+], Wennan Chang[1,3], Changlin Wan[1,3], Baskar Ramdas[4], Reuben Kapur[4], Xiongbin Lu[1*], Sha Cao[1,2,5*], Chi Zhang[1, 2*]

[1]Department of Medical and Molecular Genetics and Center for Computational Biology and Bioinformatics, [4]Department of Pediatrics, [5]Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN,46202, USA. [2]Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, Indianapolis, IN, 46202, USA. [3]Department of Electrical and Computer Engineering, Purdue University, Indianapolis, IN, 46202, USA

SUPPLEMENTARY NOTES – DERIVATION OF THE MATHEMATICAL CONDITIONS OF TRANSCRIPTOMICALLY IDENTIFIABLE CELL TYPES

A preliminary evaluation of the identifiability of cell types and cell type specific functions in tissue transcriptomics data

Gene expression signal in a tissue

We conducted the following preliminary evaluation using real and simulated data, to derive the mathematical conditions of identifiable cell types and cell type specific function at transcriptomics level.

A tissue expression profile could be thought as generated by pooling mRNAs from all of its cells, the total mRNA abundance in one tissue sample can then be considered as the following model:

$$X^i_{M\times 1} = \sum_{c\in Tissue\ i} x^{c,i}_{M\times 1} + Z^i_{M\times 1}\ (1)$$

where $X^i_{M\times 1}$ denotes the mRNA abundance of M genes in the ith tissue sample and $x^{c,i}_{M\times 1}, c \in Tissue\ i$ represents the mRNA profile of the c th cell in the tissue, $Z^i_{M\times 1}$ represents the extracellular mRNA abundance. Assume there are C^i cells in tissue i, in which C^i_k number of cell are of type $k, k = 1 \dots K,$ in tissue i, such that $\sum_{k=1\dots K} C^i_k = C^i$. Then, equation (1) can be written as:

$$X^i_{M\times 1} = \sum_{k=1}^{K}\sum_{j=1}^{C^i_k} x^{k,i}_{M\times 1,j} + Z^i_{M\times 1}\ (2)$$

where $x^{k,i}_{M\times 1,j}$ denotes the gene expression profile of the jth cell of type k in the tissue sample i. The exact population and proportion of cell type k in the sample i are C^i_k and C^i_k/C^i, respectively. Denote $P^i_k = C^i_k/C^i$.

The goal of this preliminary analysis is to derive the necessary assumptions and conditions to (1) ensure high robustness of the deconvolution analysis, i.e. robust approximation of P^i_k, and (2) maximize the resolution of deconvolution analysis, i.e., to rationally estimate the proportion of more cell or sub cell types from a tissue data. We will specifically consider (i) data quality

and impact of data normalization, (ii) the commonality and difference of transcriptomically identified and biologically defined cell types, (iii) disparity of distribution between the gene expressions in training and target tissue data, for which domain adaptation needs to be handled.

The impact of averaged count sampling and gene expression normalization

Caused by the PCR amplification and limited by the sequencing amount, denoted $\tilde{X}^i_{M\times 1}$ as the measured gene expression profile from $X^i_{M\times 1}$, which is first normalized by the observed gene expression signal through different tissue samples. Denote $|X_{M\times 1}|_{L_1}$ as the sum of all elements in $X_{M\times 1}$, i.e., the L_1 norm, and we have

$$\tilde{X}^i_{M\times 1} = \mathrm{B_i} \cdot \frac{\sum_{k=1}^{K}\sum_{j=1}^{C^i_k} x^{k,i}_{M\times 1,j} + Z^i_{M\times 1} + \epsilon_{0,i}}{\sum_{k=1}^{K}\sum_{j=1}^{C^i_k} \left|x^{k,i}_{M\times 1,j}\right|_{L1} + |Z^i_{M\times 1}|_{L1}} + \epsilon_{1,i} \quad (3)$$

, where $\mathrm{B_i}$ denotes the total observed gene expression signal in sample i, an analog of the total sequenced counts of RNA-seq data, $\epsilon_{0,i}$ and $\epsilon_{1,i}$ represent the library preparation (or sample preparation) error and measurement error respectively.

Let $\bar{x}^{k,i}_{M\times 1}$ denote the mean expression profile of the cells of cell type k in the ith tissue sample, we have

$$\tilde{X}^i_{M\times 1} = B_{\mathrm{i}} \cdot \frac{\sum_{k=1}^{K}\sum_{j=1}^{C^i_k} x^{k,i}_{M\times 1,j} + Z^i_{M\times 1} + \epsilon_0}{\sum_{k=1}^{K} C^i_k \left|\bar{x}^{k,i}_{M\times 1}\right|_{L_1} + |Z^i_{M\times 1}|_{L1}} + \epsilon_1 \quad (4)$$

Denote $\bar{B}_i = \frac{\sum_{k=1}^{K} C^i_k \bar{x}^{k,i}_{M\times 1} + |Z^i_{M\times 1}|_{L1}}{C^i} = \sum_{k=1}^{K} P^i_k \bar{x}^{k,i}_{M\times 1} + \frac{|Z^i_{M\times 1}|}{C^i}$, we further have

$$\tilde{X}^i_{M\times 1} = \frac{B_{\mathrm{i}}}{\bar{B}_i}\sum_{k=1}^{K}\frac{\sum_{j=1}^{C^i_k} x^{k,i}_{M\times 1,j}}{C^i} + \frac{B_{\mathrm{i}}}{\bar{B}_i}\frac{Z^i_{M\times 1}}{C^i} + \frac{B_{\mathrm{i}}}{\bar{B}_i}\epsilon_0 + \epsilon_1 \quad (5)$$

Equation (5) characterizes the relationships among the gene expression signal from each single cell, sample wise variations in cell type proportion, total observed gene expression signal, extracellular mRNA and errors on the observed tissue level gene expression. Based on (5), we will investigate the impact of different assumptions on estimating P^i_k, for the deconvolution of N tissue samples $\tilde{X}_{M\times N} = \{\tilde{X}^i_{M\times 1} | i = 1 \dots N\}$.

Historically, deconvolution assumes $\tilde{X}_{M\times N}$ as the following non-negative product form:

$$\tilde{X}_{M_0\times N} = \tilde{S}_{M_0\times K_0} \cdot \tilde{P}_{K\times N} + E, \tilde{S}_{M_0\times K_0} \geq 0, \tilde{P}_{K_0\times N} \geq 0 \quad (6)$$

,where $\tilde{X}_{M_0\times N}$ represents the observed gene expression matrix of M_0 selected genes in N tissue samples, and $\tilde{S}_{M_0\times K_0}$, usually unknown, denotes the gene expression signature matrix, and $\tilde{P}_{K_0\times N}$ denotes the proportion matrices of K_0 cell types, to be estimated. To estimate $\tilde{P}_{K_0\times N}$, the loss function of a deconvolution problem is usually written as

$$\min\left(\left\|\tilde{X}_{M_0\times N} - \tilde{S}_{M_0\times K_0} \cdot \tilde{P}_{K\times N}\right\|_{L2} + penalty \mid constraint\right) \quad (7)$$

, where $\|X\|_{L_2}$ denotes L_2 norm of X, i.e., sum of square of all elements, and the $penalty$ and $constraint$ vary by methods. Let $\tilde{P}_{K_0\times N}$ be a solution to the optimization problem in (7), then, $E\left(\tilde{X}^i_{M_0\times 1} - \tilde{S}_{M_0\times K_0} \cdot \tilde{P}^i_{K_0\times 1}\right) \approx 0$. Successful deconvolution using formulation (7) needs

to ensure $\tilde{P}_{K_0\times N}$ to be close to the true proportion matrix, $P_{K_0\times N}$, meaning either their absolute deviance is small, or loosely, their Pearson correlation is high.

Note

$$E\left(\tilde{X}^i_{M_0\times 1}-\tilde{S}_{M_0\times K_0}\cdot\ \tilde{P}^i_{K_0\times 1}\right)$$

$$=E\left(\frac{B_{\mathrm{i}}}{\bar{B}_i}\sum_{k=1}^{K}\frac{\sum_{j=1}^{C^i_k}x^{k,i}_{M_0\times 1,j}}{C^i}+\frac{B_{\mathrm{i}}}{\bar{B}_i}\frac{Z^i_{M_0\times 1}}{C^i}+\frac{B_{\mathrm{i}}}{\bar{B}_i}\epsilon_0+\epsilon_1-\tilde{S}_{M_0\times K_0}\cdot\ \tilde{P}^i_{K_0\times 1}\right)\quad(8)$$

It is reasonable to assume i.i.d B_{i}, ϵ_0 and ϵ_1 through different samples. Though $E(\epsilon_0)=0$ and $E(\epsilon_1)=0$ may not hold due to non-negative constraints, we assume $E(\epsilon_0)$ and $E(\epsilon_1)$ be substantially small and could be thus negligible. Later, the impact of error, ϵ_0,ϵ_1, and data scale will be further discussed. Note that since it is usually unknown what cell types are existent in the tissue, the cell types targeted in $\tilde{S}_{M_0\times K_0}$ may either contain cell types not existent or miss certain cell types. Denote $\tilde{P}^i_{K_0\times 1}$ as the ith row of $\tilde{P}_{K_0\times N}$, we have

$$E\left(\tilde{X}^i_{M_0\times 1}-\tilde{S}_{M_0\times K_0}\cdot\ \tilde{P}^i_{K_0\times 1}\right)$$

$$\approx\sum_{k=1}^{K_1}E\left(\frac{B_{\mathrm{i}}}{\bar{B}_i}\bar{x}^{k,i}_{M_0\times 1}P^i_k-\tilde{S}^k_{M_0\times 1}\cdot\tilde{P}^i_k\right)+E\left(\frac{B_{\mathrm{i}}}{\bar{B}_i}\frac{Z^i_{M_0\times 1}}{C^i}\right)$$

$$+\sum_{k=K_1+1}^{K_2}E\left(\frac{B_{\mathrm{i}}}{\bar{B}_i}x^{k,i}_{M_0\times 1,j}P^i_k\right)-\sum_{k=K_2+1}^{K_3}\tilde{S}^k_{M_0\times 1}\cdot\tilde{P}^i_k\quad(9)$$

where $\tilde{P}^i_k$ denotes the kth element in $\tilde{P}^i_{K_0\times 1}$, $\tilde{S}^k_{M_0\times 1}$ denotes the kth column of $\tilde{S}_{M_0\times K_0}$. Here, we divide the cell types into three subsets, $\{1\dots K_1\},\{K_1+1,\dots,K_2\}$ and $\{K_2+1,\dots,K_3\}$, and the first subsets contains cell types that are indeed present in the tissue i, and are also concidered in signature matrix $\tilde{S}_{M_0\times K_0}$; while the second subset contains cell types not considered in matrix $\tilde{S}_{M_0\times K_0}$ but are in fact present in tissue sample i; and the third subset contains cell types considered in matrix $\tilde{S}_{M_0\times K_0}$, but are in fact not present in tissue i.

The decomposition in (9) suggests that in order for $\tilde{P}^i_k$ to be an unbiased estimator of P^i_k, these conditions should hold: (i) $E\left(\frac{B_{\mathrm{i}}}{\bar{B}_i}\frac{Z^i_{M_0\times 1}}{C^i}\right)$, $\sum_{k=K_1+1}^{K_2}E\left(\frac{B_{\mathrm{i}}}{\bar{B}_i}x^{k,i}_{M_0\times 1,j}P^i_k\right)$ and $\sum_{k=K_2+1}^{K_3}\tilde{S}^k_{M_0\times 1}\cdot\tilde{P}^i_k$ should be all zero or orthogonal to $\sum_{k=1}^{K_1}E\left(\frac{B_{\mathrm{i}}}{\bar{B}_i}x^{k,i}_{M_0\times 1,j}P^i_k\right)$ and $\sum_{k=1}^{K_1}\tilde{S}^k_{M_0\times 1}\cdot\tilde{P}^i_k$, (ii) $\tilde{S}^k_{M_0\times 1}$ should be close to $E(\frac{B_{\mathrm{i}}}{\bar{B}_i}x^{k,i}_{M_0\times 1,j})$ and (iii) $\tilde{S}^k_{M_0\times 1}$ (and $E(x^{k,i}_{M_0\times 1,j})$) should be orthogonal for different k to avoid the impact of co-linearity in predicting $\tilde{P}^i_k$.

Impact of extracellular mRNA and mRNA from untargeted cells

Considering the non-negative property of $\frac{B_{\mathrm{i}}}{\bar{B}_i}x^{k,i}_{M_0\times 1,j}$, P^i_k, $\tilde{S}^k_{M_0\times 1}$ and $\tilde{P}^i_{K_0\times 1}$, a good selection of marker genes and determination of cell types could ensure $E\left(\frac{B_{\mathrm{i}}}{\bar{B}_i}\frac{Z^i_{M_0\times 1}}{C^i}\right)=0$ and other orthogonality conditions needed in (i). Even if some cell types are not covered in $\tilde{S}_{M_0\times K_0}$, if the marker genes in M_0 were properly selected, the condition (i) can still hold as

$\sum_{k=K_1+1}^{K_2} E\left(\frac{B_\mathrm{i}}{\bar{B}_i} x_{M_0\times 1,j}^{k,i} P_k^i\right) = 0$. Due to marker genes were always selected based on the knowledge of the predefined cell types, On the other hand, if a certain cell type t in $\tilde{S}_{M_0\times K_0}$ s in fact not present in the tissue sample, the impact of the cell type can only be eliminated if $\tilde{S}_{M_0\times 1}^t \cdot \tilde{P}_t^i$ does not increase the goodness of fitting of $\tilde{X}_{M_0\times 1}^i$ under the non-negative assumption, i.e. the inner product of $\sum_{k=1}^{K_1} E\left(\frac{B_\mathrm{i}}{\bar{B}_i} x_{M_0\times 1,j}^{k,i} P_k^i - \tilde{S}_{M_0\times 1}^k \cdot \tilde{P}_k^i\right)$ and $\tilde{S}_{M_0\times 1}^t$ is not larger than 0. However, this condition is hard to be controlled in real data, and what we could do is to make sure $\tilde{S}_{M_0\times 1}^t$ is orthogonal to $\sum_{k=1}^{K_1} E\left(\frac{B_\mathrm{i}}{\bar{B}_i} x_{M_0\times 1,j}^{k,i} P_k^i\right)$ and each $\tilde{S}_{M_0\times 1}^k, k = 1 \ldots K_1$, as much as possible. Practically, the genes of which $\tilde{S}_{M_0\times 1}^t \neq 0$ should be none or lowly expressed in the tissue.

Condition I: Signature genes in M_0 should be properly selected such that enable $E\left(\frac{B_\mathrm{i}}{\bar{B}_i} \frac{Z_{M_0\times 1}^i}{C^i}\right) = 0$, $\sum_{k=K_1+1}^{K_2} E\left(\frac{B_\mathrm{i}}{\bar{B}_i} x_{M_0\times 1,j}^{k,i} P_k^i\right) = 0$, and additionally, genes of which $\tilde{S}_{M_0\times 1}^t \neq 0, t \in \{K_2+1, \ldots, K_3\}$ should be none or lowly expressed in the tissue data.

Under ***Condition I***, and if none of the cell proportions is linear dependent, then the gene expression profile of the properly selected M_0 markers genes $\tilde{X}_{M_0\times N}$ is of rank K_1, with basis in the column/row vectors of $\tilde{S}_{M_0\times K_1}/\tilde{P}_{K_1\times N}, K_1 < M_0, K_1 < N$.

Impact of the co-linearity of the gene expression signature of different cell types.

The impact of the variations of $\bar{B}_i$, B_I and $E\left(x_{M_0\times 1,j}^{k,i}\right)$ of same cell type k through different samples will be discussed in further sections. Here, for simplicity, we assume constant $B \triangleq \frac{B_\mathrm{i}}{\bar{B}_i}$ and $E\left(x_{M_0\times 1,j}^{k,i}\right) \triangleq \hat{S}_{M_0\times 1}^k$ through different samples and that ***Condition I*** holds. Consider the case of $K_1 = 2$, i.e., $E\left(\tilde{X}_{M_0\times N}\right) = B(\hat{S}_{M_0\times 1}^1, \hat{S}_{M_0\times 1}^2) \cdot (P_{2\times N})^T$. Since predicted cell proportions of deconvolution analysis will be mainly used to conduct association analysis with other tissue or molecular features. Hence, instead of absolute deviance, we use the Pearson Correlation Coefficients (PCC) between the real and predicted proportion of each cell type through all tissue samples, $P_k^{i=1\ldots N}$ and $\tilde{P}_k^{i=1\ldots N}$, to evaluate their similarity. In supervised deconvolution, the signature matrix is uslaly pre-given or predicted using independent data. Denote the true and pre-given/predicted cell type specific gene expression signatures as, $\hat{S}_{M_0\times 1}^1, \hat{S}_{M_0\times 1}^2$ and $\tilde{S}_{M_0\times 1}^1, \tilde{S}_{M_0\times 1}^2$, we here examine the impact of signature gene expression co-linearity on the PCC of the rows of the true and predicted proportion matrices, $P_{2\times N}$ and $\tilde{P}_{2\times N}$. Here, $\tilde{P}_{2\times N}$ corresponds to $\tilde{S}_{M_0\times 2}$, and are the solution to minimizing $\left\|\tilde{X}_{M_0\times N} - \tilde{S}_{M_0\times 2} \cdot \tilde{P}_{2\times N}\right\|$.

Denote $D_{2\times 2}$ as the matrix of the cosine similarity between $\hat{S}_{M_0\times 1}^k$ and $\tilde{S}_{M_0\times 1}^k$ for $k = 1, 2$, i.e.

$$D_{2\times 2}(i,j) = \frac{\hat{S}^i_{M_0\times 1}\cdot \tilde{S}^j_{M_0\times 1}}{|\hat{S}^i_{M_0\times 1}|\cdot|\tilde{S}^j_{M_0\times 1}|}$$

, and

$$\widehat{D}_{12} = \frac{\hat{S}^1_{M_0\times 1}\cdot \hat{S}^2_{M_0\times 1}}{|\hat{S}^1_{M_0\times 1}|\cdot|\hat{S}^2_{M_0\times 1}|}$$

(1) Suppose $D_{2\times 2}(1,1)$ and $D_{2\times 2}(2,2)$ are close to 1 and $var(\tilde{X}^i_{M_0\times 1})$ is small, then, $cor\left(\tilde{P}^{i=1\dots N}_k, P^{i=1\dots N}_k\right)\approx 1, k=1,2$. Large $D_{2\times 2}(1,1)$ and $D_{2\times 2}(2,2)$ suggest high linear dependency of the true and pre-given/predicted marker genes, and small $var(\tilde{X}^i_{M_0\times 1})$ suggest lowly varied signature genes in M_0 across patients. However, in the analysis of real data set, it is almost impossible to make these conditions hold (see further real scRNA-seq and tissue data analysis).

(2) Suppose $\hat{S}^1_{M_0\times 1}\perp \hat{S}^2_{M_0\times 1}$. If $cor\left(\hat{S}^1_{M_0\times 1},\tilde{S}^2_{M_0\times 1}\right)<0$ and $cor\left(\hat{S}^2_{M_0\times 1},\tilde{S}^1_{M_0\times 1}\right)<0$, under the non-negative condition, $\tilde{P}^1_{M_0\times 1}$ and $\tilde{P}^2_{M_0\times 1}$ will have zero loadings on $\hat{S}^2_{M_0\times 1}$ and $\hat{S}^1_{M_0\times 1}$, respectively. In addition, if $cor\left(\hat{S}^1_{M_0\times 1},\tilde{S}^1_{M_0\times 1}\right)>0$ and $cor\left(\hat{S}^2_{M_0\times 1},\tilde{S}^2_{M_0\times 1}\right)>0$, $\tilde{P}^1_{M_0\times 1}$ and $\tilde{P}^2_{M_0\times 1}$ will be exactly the projection of $\tilde{X}_{M_0\times N}$ on $\hat{S}^1_{M_0\times 1}$ and $\hat{S}^2_{M_0\times 1}$ respectively, so are $\hat{P}^1_{M_0\times 1}$ and $\hat{P}^2_{M_0\times 1}$. Hence, $cor\left(\tilde{P}^{i=1\dots N}_1, P^{i=1\dots N}_1\right)=1$ and $cor\left(\tilde{P}^{i=1\dots N}_2, P^{i=1\dots N}_2\right)=1$.

(3) Suppose $\hat{S}^1_{M_0\times 1}$ is not orthogonal to $\hat{S}^2_{M_0\times 1}$. If there are gene sets $M_1\subset M_0$ and $M_2\subset M_0$, $M_1\cap M_2=\emptyset$, such that $\hat{S}^2_{M_1\times 1}=0$ and $\hat{S}^1_{M_2\times 1}=0$, we could replace M_0 by $\{M_1,M_2\}$ to make $\hat{S}^1_{M_0\times 1}$ and $\hat{S}^2_{M_0\times 1}$.= are orthogonal. Then (2) holds.

(4) Suppose $\hat{S}^1_{M_0\times 1}$ is not orthogonal to $\hat{S}^2_{M_0\times 1}$, and there are no "uniquely" expressed genes in $\hat{S}^1_{M_0\times 1}$ and $\hat{S}^2_{M_0\times 1}$ as in (3). The solution to (6) is not unique. And the bias of $\tilde{P}^{i=1\dots N}_1$ and $\tilde{P}^{i=1\dots N}_2$ from true proportions depend on the level of linear dependency of $\hat{S}^1_{M_0\times 1}$ and $\hat{S}^2_{M_0\times 1}$. Here, completely unsupervised deconvolution will be problematic. However, if $\left(\tilde{S}^1_{M_0\times 1},\tilde{S}^2_{M_0\times 1}\right)$ could be derived from another independent data source such that they are close to $\hat{S}^1_{M_0\times 1}$ and $\hat{S}^2_{M_0\times 1}$, $\tilde{P}^{i=1\dots N}_1$ and $\tilde{P}^{i=1\dots N}_2$ estimated using pre-fixed $\tilde{S}^1_{M_0\times 1},\tilde{S}^2_{M_0\times 1}$ will be highly positively associated with the true proportions..

Condition II.1 (non-negative regression formulation). For K_1 cell types that both present in the tissue samples and are considered in the $\tilde{S}_{M_0\times K_1}$ matrix, if $\hat{S}^i_{M_0\times 1}\perp \hat{S}^j_{M_0\times 1}|i,j\in\{1\dots K_1\}, i\neq j$, and if for any $\hat{S}^i_{M_0\times 1}$, the best non-negative regression of $\hat{S}^i_{M_0\times 1}=\sum_{j=\{1\dots K_1\}\backslash\{i\}}\beta_j\hat{S}^j_{M_0\times 1}$ is $\beta_j=0, j=\{1\dots K_1\}\backslash\{i\}$, i.e. $\hat{S}^i_{M_0\times 1}$ is not positively associate with any non-negative linear combination of $\hat{S}^j_{M_0\times 1}, j=\{1\dots K_1\}\backslash\{i\}$, and $cor\left(\hat{S}^k_{M_0\times 1},\tilde{S}^k_{M_0\times 1}\right)>$

$0, k = 1 \dots K_1$ and $cor\left(\hat{S}^i_{M_0\times 1}, \tilde{S}^j_{M_0\times 1}\right) < 0, i \neq j$, then $cor\left(\tilde{P}^{i=1\dots N}_k, P^{i=1\dots N}_k\right) = 1, k = 1 \dots K_1$.

It is noteworthy that some methods, including CIBERSORT, solves a regular matrix decomposition in minimizing $\left\|\tilde{X}_{M_0\times N} - \tilde{S}_{M_0\times K_1} \cdot \tilde{P}_{K_1\times N}\right\|$, and then force all negative elements in $\tilde{P}_{K_1\times N}$ to be zero to ensure its non-negativity. Under such a formulation, $\tilde{P}_{K_1\times N}$ will suffer severely from bias caused by co-linearity between cell type specific gene expression signatures (1).

Condition II.2 (general regression formulation). For the K_1 cell types that both present in the tissue samples and are considered in the $\tilde{S}_{M_0\times K_1}$ matrix, if $\hat{S}^i_{M_0\times 1} \perp \hat{S}^j_{M_0\times 1} | i, j \in \{1 \dots K_1\}, i \neq j$, and for any i, $cor\left(\hat{S}^i_{M_0\times 1}, \tilde{S}^i_{M_0\times 1}\right) > 0, \forall i$ and $\hat{S}^i_{M_0\times 1} \perp \hat{S}^j_{M_0\times 1}, \forall\, i \neq j$, then $cor\left(\tilde{P}^{i=1\dots N}_k, P^{i=1\dots N}_k\right) = 1, k = 1 \dots K_1$.

Condition II.3 (necessary but not sufficient condition for un-orthogonal cell type signature expression). Suppose $\exists\, i, j \in \{1 \dots K_1\}$, s.t. $\hat{S}^i_{M_0\times 1}$ is not orthogonal to $\hat{S}^j_{M_0\times 1}$, then the necessary but not sufficient condition for $cor\left(\tilde{P}^{i=1\dots N}_k, P^{i=1\dots N}_k\right) = 1, k = 1 \dots K_1$ is that $\tilde{P}_{K_1\times N}$ provides a good non-negative bases for $\tilde{X}_{M_0\times N}$, meaning $\tilde{P}_{K_1\times N}$ and its corresponding $\tilde{S}_{M_0\times K_1}$ is (close to) an optimal solution to NMF of $\tilde{X}_{M_0\times N}$.

It is noteworthy that if the ***Condition II.1*** or ***Condition II.2*** holds, and the matrix rank of $\tilde{X}_{M_0\times N}$ is K_1, **the predicted** $\tilde{P}^{i=1\dots N}_k$ is the unique solution for the NMF of $\tilde{X}_{M_0\times N}$.

Impact of the co-linearity of the cell type proportions in supervised formulation.
Co-linearity of the cell type proportions is a result of the immune defense mechanisms, and it is not controllable through selection of markers genes. The co-linearity of cell proportions will largely affect the predicted proportions since similar (co-linear) gene expression signals from two cell types may be inferred as the signal of one cell type, the proportion of one of the two cell types will be overly estimated while the other will be under predicted.
When the ***Condition II.1*** or ***Condition II.2*** holds, performance of supervised deconvolution method won't be affected by the co-linearity between cell proportions as the solution is unique and stable. Still for the $K_0 = 2$ case, when $\hat{S}^1_{M_0\times 1}$ not orthogonal to $\hat{S}^2_{M_0\times 1}$, the dependency between cell proportions $\tilde{P}^{i=1\dots N}_1$ and $\tilde{P}^{i=1\dots N}_2$ will shrink the two dimensional solution space of $\bar{P}_{2\times N}$ in the NMF $(\hat{S}^1_{M_0\times 1}, \hat{S}^2_{M_0\times 1}) \cdot (P_{2\times N})^T$ towards one dimension. Hence the co-linearity between cell proportions will lead to easier hitting of $\tilde{P}^{i=1\dots N}_1$ and $\tilde{P}^{i=1\dots N}_2$ to the NMF solution space of $(\hat{S}^1_{M_0\times 1}, \hat{S}^2_{M_0\times 1}) \cdot (P_{2\times N})^T$. However, it will be more likely to cause low correlation between $(\tilde{P}^{i=1\dots N}_1, \tilde{P}^{i=1\dots N}_2)$ and $P_{2\times N}$ due to the shrinkage of cell type specific proportion part will also decrease the loss of the inaccurate prediction of this part. The extreme shrinkage will be achieved by linear dependent $\tilde{P}^{i=1\dots N}_1$ and $\tilde{P}^{i=1\dots N}_2$, which will cause $(\hat{S}^1_{M_0\times 1}, \hat{S}^2_{M_0\times 1}) \cdot (P_{2\times N})^T$ be a rank one matrix and result in a correct prediction of one cell

type and an arbitrary prediction of the other cell type. Hence, if ***Condition II.1*** or ***Condition II.2*** does not hold, the dependency between $\tilde{P}_1^{i=1\dots N}$ and $\tilde{P}_2^{i=1\dots N}$ can increase the prediction of the overall. In some cases, the ratio between two dependent cell types more affect the biological characteristics of the tissue, which cannot be reliably inferred by supervised method.

Impact of the co-linearity of the gene expression signature and cell type proportions in unsupervised formulation

Since no sparse and independence constraint can be made to rows of $\tilde{P}_{K_1\times N}$, the NMF of $\tilde{X}_{M_0\times N}$ has unique solution only if $\hat{S}^i_{M_0\times 1} \perp \hat{S}^j_{M_0\times 1} | i,j \in \{1 \dots K_1\}, i \neq j$ ((2)). On the other hand, if there exists $i,j \in \{1 \dots K_1\}$ s.t. $\hat{S}^i_{M_0\times 1}$ is not orthogonal to $\hat{S}^j_{M_0\times 1}$, the prediction of $\tilde{P}_{K_1\times N}$ could be both biased and non-unique. A possible improvement of the NMF based formulation is by additionally constraining the signature matrix. Denote the $\bar{S}_{M_0\times K_1}$ and $\bar{P}_{K_1\times N}$ as the fitted signature and proportion matrix. A straightforward idea is to minimize

$$\left\|\tilde{X}_{M_0\times N} - \bar{S}_{M_0\times K_1}\cdot \bar{P}_{K_1\times N}\right\| - \lambda \cdot trace(\bar{S}^T_{M_0\times K_1}\cdot \hat{S}_{M_0\times K_1})$$

, where $trace\left(\bar{S}^T_{M_0\times K_1}\cdot \hat{S}^j_{M_0\times K_1}\right)$ characterizes the similarity between $\bar{S}_{M_0\times K_1}$ and $\hat{S}_{M_0\times K_1}$, and λ is the hyperparameter. Since $\hat{S}_{M_0\times K_1}$ is unknown, we could use $\tilde{S}_{M_0\times K_1}$ as a certain approximation of $\hat{S}_{M_0\times K_1}$.

Condition III.1 (Uniqueness of solution). For the deconvolution of a tissue data generated by $\tilde{X}_{M_0\times N} = \hat{S}_{M_0\times K_1}\cdot P_{K_1\times N}$, the NMF problem $\left\|\tilde{X}_{M_0\times N} - \bar{S}_{M_0\times K_1}\cdot \bar{P}_{K_1\times N}\right\|$ has a unique solution $\hat{S}_{M_0\times K_1}$ and $P_{K_1\times N}$ only if $\hat{S}^i_{M_0\times 1} \perp \hat{S}^j_{M_0\times 1} | i,j \in \{1 \dots K_1\}, i \neq j$.

Condition III.2 (Signature matrix constraint). A sufficient condition for the constraint NMF problem $\left\|\hat{S}_{M_0\times K_1}\cdot P_{K_1\times N} - \bar{S}_{M_0\times K_1}\cdot \bar{P}_{K_1\times N}\right\| - \lambda \cdot trace(\bar{S}^T_{M_0\times K_1}\cdot \tilde{S}_{M_0\times K_1})$ has a unique solution $\hat{S}_{M_0\times K_1}$ if $\bar{S}_{M_0\times K_1} = \hat{S}_{M_0\times K_1}$ minimizes $trace(\bar{S}^T_{M_0\times K_1}\cdot \tilde{S}_{M_0\times K_1})$.

Condition III.2 is suggesting that among the multiple solutions that minimize $\left\|\tilde{X}_{M_0\times N} - \bar{S}_{M_0\times K_1}\cdot \bar{P}_{K_1\times N}\right\|$, we should select the one that most resembles the prior $\tilde{S}_{M_0\times K_1}$.

Co-linearity of cell proportions in real data

Our preliminary analysis on TCGA data suggested correlation among selected immune and stromal cells can be as high as 0.68 while the correlation between T cell and macrophage goes up to 0.94 and the correlation between fibroblast and endothelial cells can be as high as 0.92 (**Figure 1 of Supplementary Notes**). In addition, our simulation based study using tissue data artificially constructed from scRNA-seq data revealed that the co-linearity among cell types can cause a dramatic drop in prediction accuracy in the case of (1) supervised methods with a slightly biased signature matrix and (2) unsupervised NMF methods with no prior on $\bar{S}_{M_0\times K_0}$ (**Supp Fig 4**).

These real and simulated data based analyses elucidated the common co-linearity of cell components in complex tissue. Based on these observations and previous derivations, to ensure a high $cor\left(P_{1\times N}^{k}, \tilde{P}_{1\times N}^{k}\right)$, supervised methods should (i) properly select marker genes that satisfy ***Condition I*** and ***Condition II.1*** or ***Condition II.2*** or (ii) ensure extremely high consistency between $\hat{S}_{M_0\times K_1}$ and $\tilde{S}_{M_0\times K_1}$; unsupervised methods should (iii) properly select marker genes that satisfy ***Condition I*** and ***Condition III.1*** or (iv) properly select marker genes that satisfy ***Condition I*** and properly constrain $\bar{S}_{M_0\times K_1}$ such that ***Condition III.2*** will hold.

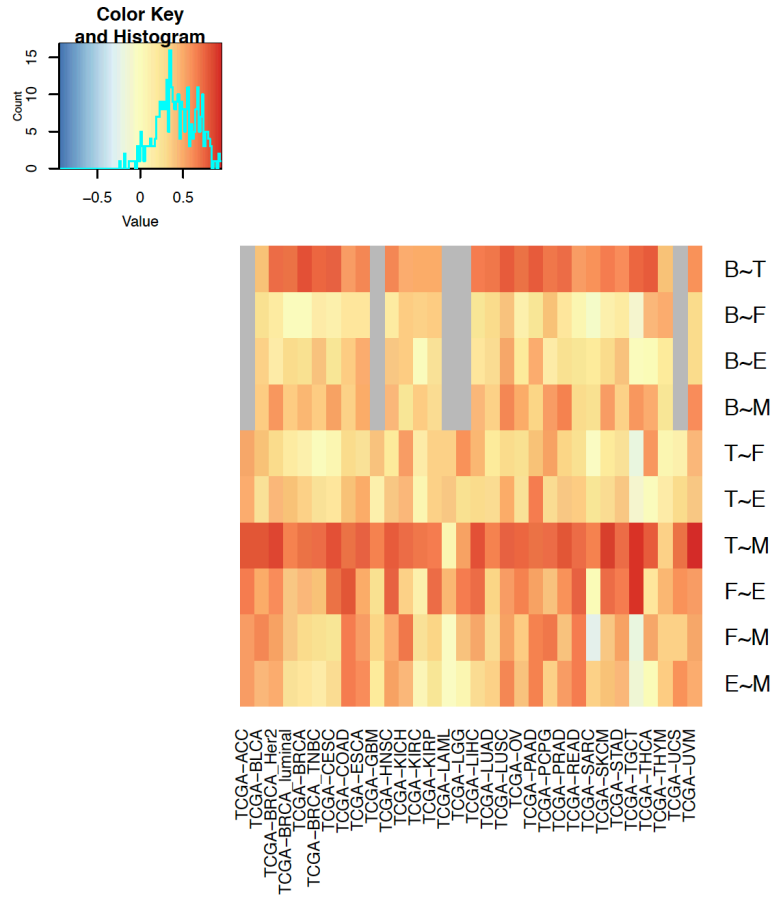


Figure 1 of Supplementary Notes. Correlation between the level of B cell (B), T cell (T), Endothelial (E), and Fibroblast cell (F) in TCGA cancer types

Difference between cell type specific gene expression profile in tissue data of different tissue data types, experimental batches and platforms.

The purpose of tissue data deconvolution analysis is to maximally utilize limited training information to predict cell types and their proportions in the tissue data largely available and can be generated with a low cost in broad research fields. To the best of our knowledge, only less than five of the tissue data sets in the public domain are with matched training data collected from the same tissue or disease condition and the training data are measured with similar experimental platform and batches so the bias between tissue and training data can be ignored. Hence

A key challenge of the supervised deconvolution method is to fixate a $\tilde{S}^k_{M_0\times 1}$ as close to $\hat{S}^k_{M_0\times 1}$ as possible so that P^i_k can be accurately estimated by $\tilde{P}^i_k$. However, $\hat{S}^k_{M_0\times 1}$ is usually not attainable, and it varies with experimental platform and batches. We first evaluate if $\hat{S}^k_{M_0\times 1}$ forms a certain distribution through different data sets. Intuitively, if $\hat{S}^k_{M_0\times 1}$ does not varied a lot through data sets, under the ***Condition I***, a constant $\hat{S}^k_{M_0\times 1}$ close enough to $\hat{S}^k_{M_0\times 1}$ can enable a good estimation of P^i_k. On the other hand, if $\hat{S}^k_{M_0\times 1}$ tends to be less consistent, a certain domain adaptation approach will be needed to ensure either ***Condition II.1*** or ***Condition II.2***, or the transformed $\tilde{S}^k_{M_0\times 1}$ be close enough to the varied $\hat{S}^k_{M_0\times 1}$.

We evaluated the distribution of $\hat{S}^k_{M_0\times 1}$ of T-, B-, macrophage, endothelial and fibroblast cells through five sets of single cell data sets of cancer microenvironment. Our analysis clearly suggested $\hat{S}^k_{M_0\times 1}$ of cell type specific marker genes varied a lot through different single cell data set (**supp Fig 1**). We selected five known cell type uniquely expressed marker genes of three major immune and stromal cell types namely T cell (CD2, CD3D, CD3E, CD3G), B cell (CD19, CD22, CD79A, CD79B), and fibroblast cells (COL1A1, COL1A2, COL3A1, COL5A1). We evaluated the gene expression profile of these genes in 53 of human normal colon, inflammatory colorectal disease, benign colorectal tumor and colorectal cancer data sets. Among the data sets, we first confirmed the expression profile of a subset of the selected marker genes of each cell type form a significant rank-1 matrix in the data set, i.e., the $\hat{S}^k_{M_0\times 1}$ of the cell type uniquely expressed marker genes can be well estimated by the first column base of each rank-1 matrix. Our analysis also suggested that on average, in more than half of the 53 data sets, only subset of the tested marker genes form rank-1 matrix. Our analysis revealed the variation of cell type specific marker genes among different data set of a same tissue type and measured on same experimental platform (**supp Fig 12**).

These preliminary data analyses exclude the possibility of deriving a common signature matrix for supervised deconvolution or satisfying **Condition III.2**. Also, it is impossible to derive fixed set of cell type uniquely expressed genes to satisfy **Condition II.1**, **Condition II.2** or **Condition III.1**. In sight of this, we consider a deconvolution capability that can be generally utilized for a broad set of datasets is via identifying cell types and gene markers that satisfy **Condition I, II** and **III**. Toward this goal, we first give a rigorous definition of ***transcriptomically identifiable*** cell types. Our previous mathematical derivations provide theoretical foundation of the necessary and sufficient condition for identification of ***transcriptomically identifiable*** cell types.

Definition I. Transcriptomically identifiable cell types. For a given transcriptomics data set $X_{M\times N}$ and a deconvolution method, a cell type k defined as "transcriptomically identifiable" if its true proportion $P^k_{1\times N} \triangleq \left\{\frac{c^1_k}{c^1}, \frac{c^2_k}{c^2}, \ldots \frac{c^N_k}{c^N}\right\}$ and estimated proportion $\tilde{P}^k_{1\times N}$ are perfectly linearly correlated, i.e. $cor\left(P^k_{1\times N}, \tilde{P}^k_{1\times N}\right) = 1$

Theorem 1.1 For a supervised method, the K_1 cell types in formula (5) are ***transcriptomically identifiable*** if and only if the ***Condition I*** and ***Condition II.1*** *or* ***Condition II.2*** strictly hold, and the $E(\epsilon_0)$ and $E(\epsilon_1)$ can be ignored.
If $E(\epsilon_0)$ and $E(\epsilon_1)$ can be ignored when N is large, ***Condition I*** and ***Condition II.1*** *or* ***Condition II.2*** *ensure* $cor\left(\tilde{P}_k^{i=1\dots N}, P_k^{i=1\dots N}\right) = 1, k = 1 \dots K_1$. ■

Theorem 1.2 For a unsupervised NMF based method, the K_1 cell types in formula (5) are ***transcriptomically identifiable*** if and only if the ***Condition I*** and ***Condition III.1*** strictly hold, and the $E(\epsilon_0)$ and $E(\epsilon_1)$ can be ignored.
Since Condition I and Condition III.1 enable uniqueness solution of the NMF problem, we have $cor\left(P_{1\times N}^k, \bar{P}_{1\times N}^k\right) = 1$ if $E(\epsilon_0)$ and $E(\epsilon_1)$ can be ignored when N is large. ■

Corollary 1.3 If the K_1 cell types in formula (5) are ***transcriptomically identifiable*** by a supervised method, they are also ***transcriptomically identifiable*** by an unsupervised method.
Easy to show as ***Condition II.1*** *or* ***Condition II.2*** are special cases of ***Condition III.1.*** ■

Revisit the advantage of supervised deconvolution method
It is noteworthy that the ***Theorem 1.1*** and ***Theorem 1.2*** describe the general conditions that a deconvolution method should satisfy for identification of ***transcriptomically identifiable*** cell types in any given tissue data sets. Based on the ***Corollary 1.3***, a supervised deconvolution enable prediction of ***transcriptomically identifiable*** cell types can be replaced by a unsupervised method. Indeed, supervised method is less favored to be applied to a broad range of data to the fixed cell type and signature assumption. However, the supervised method has its specific advantage (1) if the pre-identified $\tilde{S}_{M_0\times K_1}$ is similar enough to $\hat{S}_{M_0\times K_1}$ and (2) when the sample size is small. In addition, for certain cell types of high interests without cell type specifically expressed marker genes, a supervised method can ensure the prediction will be made. Selecting proper marker genes that maximally ensure the ***Condition I, Condition II.1*** and ***Condition II.2*** will improve the prediction accuracy. A special case that supervised method fits is when gene expression of sorted cells collected from a small representative tissue samples and a large set of tissue samples were simultaneously measured.

Towards a practical deconvolution capability
Based on the discussion above, it is impractical to train data set specific signature matrix. However, with selecting proper marker genes, ***Conditions I, II*** and ***III*** can be satisfied to enable identifiablilty of targeted cell types. With this idea, we consider a practical deconvolution method should maximally remove the genes contradict to the ***Conditions I, II*** and ***III***. Specifically, we can derive the conditions of transcriptomically identifiable cell types that can be detected and estimated by a deconvolution method without training data set specific signature matrix. Because we have proven if a cell type is identifiable by a supervised deconvolution method, it will be always identifiable by an unsupervised method. Here we only consider the conditions of transcriptomically identifiable cell types by unsupervised method.

Theorem 2.1 (strong identifiability). For a given tissue transcriptomics data $X_{M\times N}$, a cell type k is transcriptomically identifiable in all the samples, if and only if (1) there are genes G_k

uniquely expressed by cell type k, s.t. $E(X_{G_k \times T})$ is a rank one matrix, for $\forall T$ is a subset of $\{1, \dots, N\}$.
If (1) holds, each cell type is with uniquely expressed genes, hence are orthogonal to other cell types. Hence $\hat{S}^i_{M_0 \times 1} \perp \hat{S}^j_{M_0 \times 1}$ for $\forall$ cell types i and j, the ***Conditions I*** and ***III.1*** hold. ■

Theorem 2.2 (weak identifiability). For a given tissue transcriptomics data $X_{M \times N}$, a cell type k is transcriptomically identifiable in all the samples, if and only if (1) there are genes G_k uniquely expressed by cell type k, s.t. $E(X_{G_k \times T})$ is a rank one matrix, for $\forall T$ is a subset of $\{1, \dots, N\}$, and the proportion of a cell type k' in all the samples can be partially estimated if (2) there are genes $G_{k'}$ only expressed by cell type k' and other cell types $k_1, \dots, k_n$ satisfying (1) and the matrix rank of $E(X_{G_{k'} \times T})$'s projection to the complementary space spanned by the first row bases of each $E(X_{G_{k_1} \times T}), E(X_{G_{k_2} \times T}), \dots, E(X_{G_{k_n} \times T})$ is one, for $\forall T$ is a subset of $\{1, \dots, N\}$.
If (2) holds, with identifying all the identifiable cell types satisfying (1), the cell types satisfying (2) will also satisfy (1) after projecting $X_{M \times N}$ to the complementary space of the linear space spanned by the first row base of the cell type unique genes' expression profiles of these cell types. However, the estimated proportion of the cell types satisfying (2) can be biased to a linear combination of the proportion of the proportions of cell types $k_1, \dots, k_n$, due to if (2) holds, the $\hat{S}^{k'}_{G_{k'} \times 1}$ is not orthogonal to $\hat{S}^{k_1}_{G_{k'} \times 1}, \dots, \hat{S}^{k_n}_{G_{k'} \times 1}$ that the uniqueness solution condition of the NMF problem does not hold.
On the other hand, if a cell type k is transcriptomically identifiable. ***Conditions I*** and ***III*** should be hold to any subset of the samples. Hence either (1) or (2) should be held. ■

Corollary 2.3 (identifiability in a subset of samples). For a given tissue transcriptomics data $X_{M \times N}$, a cell type k is transcriptomically identifiable in a subset of samples, denoted as $\aleph$, if and only if (1) there are genes G_k uniquely expressed by cell type k, s.t. $E(X_{G_k \times T})$ is a rank one matrix, for $\forall T$ is a subset of $\aleph$, where $\aleph$ is a subset of $\{1, \dots, N\}$, and the proportion of a cell type k' in $\aleph$ can be partially estimated if (2) there are genes $G_{k'}$ only expressed by cell type k' and other cell types $k_1, \dots, k_n$ satisfy (1) and the matrix rank of $E(X_{G_{k'} \times T})$'s projection to the complementary space spanned by the top row bases of each $E(X_{G_{k_1} \times T}), E(X_{G_{k_2} \times T}), \dots, E(X_{G_{k_n} \times T})$ is first, for $\forall T$ is a subset of $\aleph$, where $\aleph$ is a subset of $\{1, \dots, N\}$.
By the proof of ***Theorem 2.1*** and ***Theorem 2.2.*** ■

Corollary 2.2 generalizes the identifiability condition to a subset of samples. Practically, the gene expression profile of some samples can be biasedly measured, which cause sample wise variations in cell type specific gene expressions. Under this circumstance, gene markers of the identifiable cell types satisfy the rank-1 condition (***Theorem 2.1***) in a subset of samples, which

can be still targeted by identifying rank-1 submatrix, which is a local low rank matrix identification problem.

With the derivation of ***Theorem 2.1, 2.2*** and the ***Corollary 2.3***, we are motivated to develop SSMD pipeline, in which we first identify the genes enrich to the preidentified knowledge of cell type specific markers and form a rank-1 submatrix in any subset of samples of a given tissue data. A further assessment of correct markers and their matrix ranks is set to ensure a correct NMF analysis. Then motivated by the ***Condition III*** and considering some of the identified cell type specifically expressed genes can also have slight expression in other cell types, we utilize a constraint matrix to increase the NMF accuracy.

Method evaluation (deriving E-score)

Condition II.3 provides an intuition for method evaluation. Due to both $\hat{S}_{M_0\times K_1}$ and K_1 are unknown in the analysis of a real data, we only derive necessary conditions for the detection of transcriptomically identifiable cell types. Denote $\tilde{X}_{M_0\times N}$ as the observed gene expression profile and $\hat{P}_k^{i=1\dots N}, k = 1\dots K$ as the proportion of K cell types predicted by a deconvolution method (can be either a supervised or unsupervised method). Denote $\tilde{X}_{M_0\times N}[g,]$ as the gene expression profile of gene g and $CS_g \subset \{1,\dots K\}$ as the cell types expressing= assumed in a supervised method or identified in an unsupervised method.

Definition II (explanation score, ES). The explanation score of gene g is defined by $ES_g \triangleq min\frac{\sum_{i=1}^{N}\left(\tilde{X}_{M_0\times N}[g,i]-\sum_{k\in CS_g}\beta_k\hat{P}_k^i\right)^2}{\sum_{i=1}^{N}\left(\tilde{X}_{M_0\times N}[g,i]-mean\left(\tilde{X}_{M_0\times N}[g,]\right)\right)^2} \mid \beta_k \geq 0.$

Theorem 3 (Necessary Condition). If the gene g is a marker gene of several transcriptomically identifiable cell types CS_g in all the samples, $\lim_{N\to\infty} ES_g = 0$.

If g is a marker gene of CS_g and the method accurately estimated cell proportions and ***Condition I*** holds, $E(\tilde{X}_{M_0\times N}[g,])$ should be spanned by the proportion of the cell types CS_g, which can be well fitted by the estimated proportions of the cell types. ■

Transcriptomically identifiable cell types vs biologically defined cell types.

In this study, we raised the definition of transcriptomically identifiable cell types. It is noteworthy such cell types are different to the common biologically defined cell types. For the example of cell types over hematopoietic and downstream immune cell lineage, cell types are defined as the cells under different differentiation or maturation stage over the lineage. However, gene expression profile of the cell types defined over the lineage can be highly similar that cannot be identified in tissue transcriptomic data. On the other hand, certain functional genes have similar expression patterns in different biologically defined cell types, such as the MHC class II genes always form a rank-1 module in tissue data. Dendric cell, macrophage and other antigen presenting immune cell types can expression MHC class II genes. If not all the cell types expressing MHC class II are identifiable and the $\hat{S}_{M_0\times 1}^i$ of the MHC class II genes

in different cell types are highly similar, then the MHC class II genes will be identified as the marker genes of transcriptomically identifiable cell type. We consider identifying such a cell type is rational due to it characterize all the cell types with a certain function. From the perspective in downstream association analysis, variations in such a cell type can be correlated with other biological and molecular features. Consequently, the biological definition of cell type more distinguishing the variation in physical characteristics while transcriptomically identifiable cell types more distinguishing the cells with distinct expression variations.

Inference of two highly similar cell or sub cell types

By ***Condition II*** and ***Condition III***, two cell types sharing similar signatures cannot be transcriptomically identifiable. However, with selection of proper marker genes, two closely related cell types can still satisfy ***Condition II*** or ***Condition III***. One example is via selecting cell type unique genes of CD4+ and CD8+ T cells, the two cells can be transcriptomically identifiable. One necessary condition of unique solution of an NMF problem is that for $\forall i, j$, the genes of $\hat{S}^i_{M_0\times 1} \neq 0$ is not a subset of the genes of $\hat{S}^j_{M_0\times 1} \neq 0$. Hence, the naïve and activated forms of one cell type are less likely to be identifiable if the naïve state is not with unique gene markers.

Different level of noises in tissue data and their impact to deconvolution analysis

Our formulation includes the following level of noises:

$$E\left(\tilde{X}^i_{M_0\times 1} - \tilde{S}_{M_0\times K_0} \cdot \tilde{P}^i_{K_0\times 1}\right)$$
$$= E\left(\frac{B_{\mathrm{i}}}{\bar{B}_i}\sum_{k=1}^{K}\frac{\sum_{j=1}^{C^i_k} x^{k,i}_{M_0\times 1,j}}{C^i} + \frac{B_{\mathrm{i}}}{\bar{B}_i}\frac{Z^i_{M_0\times 1}}{C^i} + \frac{B_{\mathrm{i}}}{\bar{B}_i}\epsilon_0 + \epsilon_1 - \tilde{S}_{M_0\times K_0} \cdot \tilde{P}^i_{K_0\times 1}\right)$$

(1)ϵ_0 and ϵ_1 represent the library preparation (or sample preparation) error and measurement error; (2) different $\bar{B}_i = \frac{\sum_{k=1}^{K} C^i_k \bar{x}^{k,i}_{M\times 1} + |Z^i_{M\times 1}|}{C^i} = \sum_{k=1}^{K} P^i_k \bar{x}^{k,i}_{M\times 1} + \frac{|Z^i_{M\times 1}|}{C^i}$ of different samples; and (3) different distribution of $x^{k,i}_{M_0\times 1,j}$ of the same type k in different samples i.

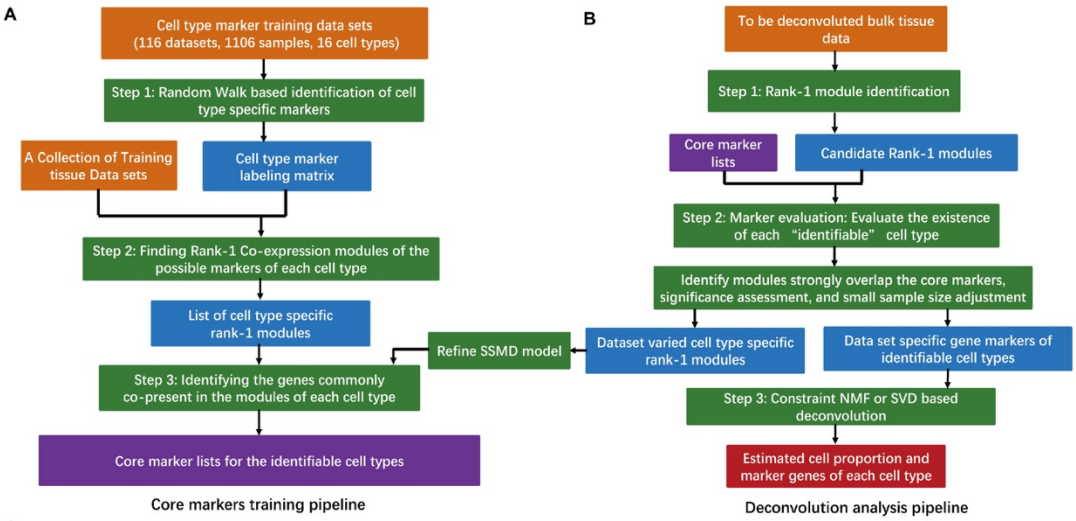
For (1), due to the observed data are non-negative, $E(\epsilon_0)$ and $E(\epsilon_1)$ are not zero. If $E(\epsilon_0)$ and $E(\epsilon_1)$ are large or varied a lot in different sample, the rank-1 pattern of cell type uniquely expressed marker genes will be less significant, that causing less transcriptomically identifiable cell types and analysis resolution. It is noteworthy the $\tilde{X}^i_{M_0\times 1}$ is of selected marker genes of identifiable cell types, which can be substantially small if the sample is will low proportion of the cell types. Hence ϵ_1 can be varied through different samples.

For (2), $\bar{B}_i$ cannot be identical for different samples i due to the variation of P^i_k. However, if $\bar{x}^{k,i}_{M\times 1}$ and $\frac{|Z^i_{M\times 1}|}{C^i}$ does not varied a lot, $\bar{B}_i$ tends to be identical. Normally, it is rational to assume a same total mRNA level $\bar{x}^{k,i}_{M\times 1}$ for different cell type k and sample j. However, in the cancer tissue, some cancer cells are with high mRNA abundance due to highly activated transcription. On the other hand, some naïve state cells may have less mRNA abundance. If this sample is with high proportion of cancer cells, the proportion of immune cell types can be

underestimated. Assuming a less variation in the second term $\frac{|Z^i_{M\times 1}|}{C^i}$ is rational due to a tissue with more cells are also with more extracellular mRNA. However, some cancer may have more autophagy or dead cells, that causing more $\frac{|Z^i_{M\times 1}|}{C^i}$. It is noteworthy the rank-1 structure of cell type specific marker genes will not change. Hence, the variations in $\bar{x}^{k,i}_{M\times 1}$ and $\frac{|Z^i_{M\times 1}|}{C^i}$ will not affect the transcriptomic identifiability but may cause a biased prediction in cell proportions. In addition, a larger $E(\frac{Z^i_{M_0\times 1}}{C^i})$ will decrease the significance of the rank-1 pattern of the cell type uniquely expressed marker genes by affecting the low rank structure of $\tilde{X}_{M_0\times N}$.

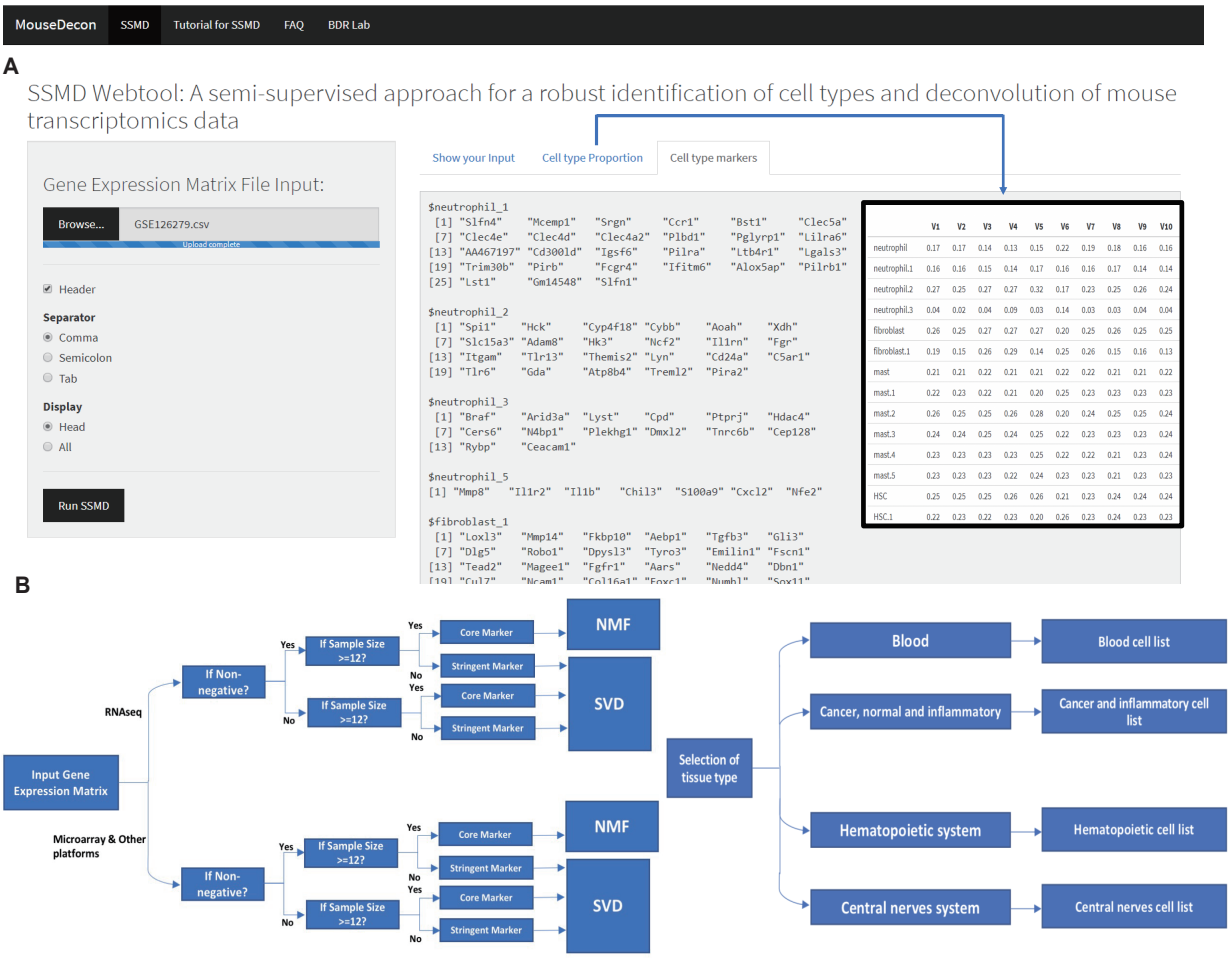
For (3), all deconvolution methods assume identical $E(x^{k,i}_{M_0\times 1,j})$ for different i. However, it is possible that the gene expression level of cell type unique markers varied through different samples. For example, T cells in different cancer tissue may have varied cytotoxic gene expression, which are commonly used as marker genes of CD8+ T cells. Variations in $E(x^{k,i}_{M_0\times 1,j})$ for different i may directly destroy the rank-1 pattern of the markers and bias the prediction. However, if the change of $E(x^{k,i}_{M_0\times 1,j})$ follows a co-activation or co-suppression pattern, i.e. the expression level of the marker genes in a unit cell co-up or down regulated in each tissue sample, instead of estimating the cell proportion, the functional associated with the marker genes can be predicted.

Supplementary Figure S1.



Supplementary Figure S1. SSMD pipeline.

Supplementary Figure S2.



Supplementary Figure S2. SSMD web server and utilization guideline.