**Text S1**

**Algorithm Pseudo code**

A matlab program of this code will be provided by the authors upon request.

**Input: (A)** $m x n$ mixed matrix $M$ and $m x k\_{max}$ reference signatures matrix $L$ with non-negative entries, where $m$ is the number of genes, $n$ is the number of samples and $k\_{max}$ is the number of initial cell-types, **(B)** A collection of $T^{'}=\left\{T\_{1}^{'}…T\_{k\_{max}}^{'}\right\}$cell-type labels where $T\_{k}^{'}$ is the label of the $k^{th}$ column in $L$, **(C)** number of majority voting runs $Q$.

**For** $v=1…Q$

(1) Initialize: $h\_{kj}$ with non-negative values $1\leq k\leq k\_{max}$, $1\leq j\leq n$

Initialize: $w\_{ik}(v)$ with $l\_{ik}$ $1\leq i\leq m$, $1\leq k\leq k\_{max}$

Scale columns of $W(v)$ to sum to one.

(2) Get $W(v)$ and $H$ using NMF, as described in Piper et al. 12:

 $H\_{kj}^{(t)}=H\_{kj}^{(t-1)}\frac{\left[\left(W^{(t-1)}\right)^{T}M\right]\_{kj}}{\left[\left(W^{(t-1)}\right)^{T}W^{(t-1)}H^{(t-1)}\right]\_{kj}+ϵ}$

 $W\_{ik}^{(t)}=H\_{ik}^{(t-1)}\frac{\left[M\left(H^{(t)}\right)^{T}\right]\_{ik}}{\left[W^{(t-1)}H^{(t)}\left(H^{(t)}\right)^{T}\right]\_{ik}+ϵ}$

where $ϵ≈10^{-9}$ is used to avoid possible division by zero and $t$ refers to
 the NMF iteration

(3) Determine $\hat{k}\_{CT}$(*v*) according to (4)

(4) Determine $\hat{G}$(*v*) and its cell-type labels as the chosen columns in $W(v)$

(5) Set $Z\_{k}\left(v\right)=1$ if the label $T\_{k}^{'}$ is chosen, $Z\_{k}\left(v\right)=0$ otherwise

**End For**

(6) Determine the final cell-type identities:

$\hat{T}=\left\{\hat{T}\_{1}, …, \hat{T}\_{k\_{CT}}\right\}$, where $T\_{k}^{'}\in \hat{T}$ if $\frac{1}{Q}\sum\_{v=1}^{Q}Z\_{k}(v)\geq threshold$

(7) Determine the final $\hat{G}$: for $i=1,…,\hat{k}\_{CT}\left(v\right)$

$$\hat{g}\_{i}=average\left\{\hat{g}\_{k}\left(v\right), v=1,…,Q, such that a label of \hat{g}\_{k}\left(v\right) is \hat{T}\_{i}\right\}$$

(8) Set $\hat{k}\_{CT}$ to the number of columns in the matrix $\hat{G}$

(9) Determine $\hat{C}$ according to Eq. (5)

**Output:** $\hat{G}$, $\hat{C}$, $\hat{k}\_{CT}$

\* To use classes the algorithm requires the following sets $Ψ\_{p}, p=1…P$, where $P$ is the number of classes and each $Ψ\_{p}$ contains the collection of labels $T\_{k}^{'}$ that are affiliated with class $p$. Note that there may be classes that contain a single label. Let$ ψ\_{p}$ be the label of class $p$, then the algoritm outputs $\hat{ψ}=\left\{\hat{ψ}\_{1}, …, \hat{ψ}\_{k\_{CT}}\right\}$ which is the estimated labels of classes as follows:

Change (6) such that:

$\hat{ψ}=\left\{\hat{ψ}\_{1}, …, \hat{ψ}\_{k\_{CT}}\right\}$, where $ψ\_{p}\in \hat{T}$ if $\frac{1}{Q}\sum\_{k:T\_{k}^{'}\in C\_{p}}^{}\sum\_{v=1}^{Q}Z\_{k}(v)\geq threshold$

Change (7) to:

Determine the final $\hat{G}$: for $i=1,…,\hat{k}\_{CT}\left(v\right)$

$\hat{g}\_{i}=average\left\{\hat{g}\_{k}\left(v\right), v=1,…,Q, such that a label of \hat{g}\_{k}\left(v\right) is in \hat{ψ}\_{i}\right\}$