BayesDeBulk: A Flexible Bayesian Algorithm for the Deconvolution of Bulk Tumor Data

Supplementary Material

Francesca Petralia[1], Azra Krek[1], Anna P. Calinawan[1], Song Feng [2], Sara Gosline[2],
Pietro Pugliese[3], Michele Ceccarelli[4], Pei Wang[11]

[1]Icahn School of Medicine at Mount Sinai, NY, USA
[2]Pacific Northwest National Laboratory, Seattle, WA, USA
[3]University of Sannio, Benevento, Italy
[4]University of Naples "Federico II", Naples, Italy

Contents

1 Gibbs Sampling

The Gibbs sampling scheme can be summarized in the following steps:

<u>Step 1</u> Sample mean parameter $\mu_{k,j}$ from a truncated Gaussian distribution:

$$\mu_{k,j} \sim N\left(\left(\frac{\sum_{i=1}^{n} M_{i,j}\pi_{i,k}}{\sigma_j} + \frac{\xi_{k,j}}{\lambda_{k,j}}\right)\left(\frac{1}{\lambda_{k,j}} + \frac{\sum_{i=1}^{n}\pi_{i,k}^2}{\sigma_j}\right)^{-1}, \left(\frac{1}{\lambda_{k,j}} + \frac{\sum_{i=1}^{n}\pi_{i,k}^2}{\sigma_j}\right)^{-1}\right) 1(\mu_{k,j} \in S_{k,j})$$

with $M_{i,j} = y_{i,j} - \sum_{s \neq k} \mu_{s,j}\pi_{i,s}$ and $S_{k,j}$ being defined as the intersection across all constraints involving $\mu_{k,j}$. This set is defined in Section 2.

<u>Step 2</u> Sample $Z_{i,k}$ from

$$Z_{i,k} \sim Binomial\left(\frac{w_k N_{[0,1]}(\pi_{i,k}; 0, 0.0001)}{w_k N_{[0,1]}(\pi_{i,k}; 0, 0.0001) + (1 - w_k)N_{[0,1]}(\pi_{i,k}; 0, \gamma_k)}\right)$$

<u>Step 3</u> Sample $\pi_{i,k}$ from a truncated univariate Gaussian defined as:

$$N_{[0,1]}\left(\sum_{j=1}^{p}\frac{M_{i,j}\mu_{k,j}}{\sigma_j}\left(\sum_{j=1}^{p}\frac{\mu_{k,j}^2}{\sigma_j} + \frac{1}{\eta_k}\right)^{-1}, \left(\sum_{j=1}^{p}\frac{\mu_{k,j}^2}{\sigma_j} + \frac{1}{\eta_k}\right)^{-1}\right)$$

with $\eta_k = \gamma_k$ if $\ell = 0$ and $\eta_k = 0.0001$ if $\ell = 1$.

<u>Step 4</u> Sample w_k from

$$Beta\left(1 + \sum_i 1(Z_{i,k} = 1), 1 + \sum_i 1(Z_{i,k} = 0)\right)$$

<u>Step 5</u> Sample γ_k from:

$$\text{Inverse-Gamma}\left(\alpha_\gamma + \sum_i 1(Z_{i,k} = 0)/2, \beta_\gamma + 0.5 \sum_{i|Z_{i,k}=0} \pi_{i,k}^2\right)$$

<u>Step 6</u> Sample σ_j from:

$$\text{Inverse-Gamma}\left(\alpha_\sigma + n/2, \beta_\sigma + 0.5\sum_{i=1}^{n}\left(y_{i,j} - \sum_{k=1}^{K}\mu_{k,j}\pi_{i,k}\right)^2\right)$$

<u>Step 7</u> Sample ρ from a uniform distribution

$$(\rho|-) \sim Uniform(0, h(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K))$$

2 Likelihood and Full Conditionals

The likelihood of BayesDeBulk model is defined as:

$$f(\boldsymbol{Y}) = \prod_{i=1}^{n}\prod_{j=1}^{p}(\sigma_j 2\pi)^{-1/2}\text{exp}\left(-\frac{1}{\sigma_j}\left(y_{i,j} - \sum_{k=1}^{K}\pi_{i,k}\mu_{k,j}\right)^2\right)$$

Full conditional of μ The full conditional of the mean expression of the j-th marker for the k-th cell-type (i.e., $\mu_{k,j}$) is derived as follows:

$$p(\mu_{k,j}|-) \propto \exp\left(-\sum_{i=1}^{n}\frac{1}{2\sigma_j}\left(y_{i,j} - \sum_{s=1}^{K}\mu_{s,j}\pi_{i,s}\right)^2\right)\exp\left(-\frac{1}{2\lambda_{k,j}}(\mu_{k,j} - \xi_{k,j})^2\right)1(S_{k,j})$$

$$p(\mu_{k,j}|-) \propto \exp\left(-\sum_{i=1}^{n}\frac{1}{2\sigma_j}\left(y_{i,j} - \mu_{k,j}\pi_{i,k} - \sum_{s\neq k}\mu_{s,j}\pi_{i,s}\right)^2\right)\exp\left(-\frac{1}{2\lambda_{k,j}}(\mu_{k,j} - \xi_{k,j})^2\right)$$

Let us define $M_{i,j} = y_{i,j} - \sum_{s\neq k}\mu_{s,j}\pi_{i,s}$

$$p(\mu_{k,j}|-) \propto \exp\left(-\sum_{i=1}^{n}\frac{1}{2\sigma_j}\left(M_{i,j} - \mu_{k,j}\pi_{i,k}\right)^2\right)\exp\left(-\frac{1}{2\lambda_{k,j}}(\mu_{k,j} - \xi_{k,j})^2\right)1(S_{k,j})$$

$$p(\mu_{k,j}|-) \propto \exp\left(-\frac{1}{\sigma_j}\left(\sum_{i=1}^{n}M_{i,j}^2 + \mu_{k,j}^2\sum_{i=1}^{n}\pi_{i,k}^2 - 2\mu_{k,j}\sum_{i=1}^{n}M_{i,j}\pi_{i,k}\right)\right)\exp\left(-\frac{1}{2\lambda_{k,j}}(\mu_{k,j} - \xi_{k,j})^2\right)1(S_{k,j})$$

$$p(\mu_{k,j}|-) \propto \exp\left(-\left(\mu_{k,j}^2\left(\frac{1}{\lambda_{k,j}} + \frac{\sum_{i=1}^{n}\pi_{i,k}^2}{\sigma_j}\right) - 2\mu_{k,j}\left(\frac{\sum_{i=1}^{n}M_{i,j}\pi_{i,k}}{\sigma_j} + \frac{\xi_{k,j}}{\lambda_{k,j}}\right)\right)\right)1(S_{k,j})$$

$$\mu_{k,j} \sim N\left(\left(\frac{\sum_{i=1}^{n}M_{i,j}\pi_{i,k}}{\sigma_j} + \frac{\xi_{k,j}}{\lambda_{k,j}}\right)\left(\frac{1}{\lambda_{k,j}} + \frac{\sum_{i=1}^{n}\pi_{i,k}^2}{\sigma_j}\right)^{-1}, \left(\frac{1}{\lambda_{k,j}} + \frac{\sum_{i=1}^{n}\pi_{i,k}^2}{\sigma_j}\right)^{-1}\right)1(S_{k,j})$$

$S_{k,j}$ is defined as the intersection across all constraints involving $\mu_{k,j}$ contained in the repulsive function. Letting $g(z) = \exp(-\tau z^{-\eta})$, $S_{k,j}$ is defined as $S_{k,j} = \cap_{s\neq k}\{x : x > \mu_{s,j}\ \&\ g(|\mu_{s,j} - x|) > \rho\}$ for $j \in I_k$ and $S_{k,j} = \cap_{s\neq k|j\in I_s}\{x : \mu_{s,j} > x\ \&\ g(|\mu_{s,j} - x|) > \rho\}$ for $j \notin I_k$.

Full conditional of π The prior distribution of $\pi_{i,k}$ is specified as follows: $\pi_{i,k} \sim w_k N_{[0,1]}(0, 0.0001) + (1 - w_k)N_{[0,1]}(0, \gamma_k)$ with $\gamma_k \sim$ Inverse-Gamma$(3, 1)$. Let $Z_{i,k} = 1$ if $\pi_{i,k} \sim N_{[0,1]}(0, 0.0001)$ and $Z_{i,k} = 0$ if $\pi_{i,k} \sim N_{[0,1]}(0, \gamma_k)$; with $Z_{i,k} \sim Binomial(w_k)$. Given this prior specification, the full conditional of $\pi_{i,k}$ is defined as

$$p(\pi_{i,k}|Z_{i,k} = \ell) \propto \exp\left(-\sum_{j=1}^{p}\frac{1}{2\sigma_j}\left(y_{i,j} - \mu_{k,j}\pi_{i,k} - \sum_{s\neq k}\mu_{s,j}\pi_{i,s}\right)^2\right)\exp\left(-\frac{1}{2\eta_k}\pi_{i,k}^2\right)$$

with $\eta_k = \gamma_k$ if $\ell = 0$ and $\eta_k = 0.0001$ otherwise. Define $T_{i,k,j} = y_{i,j} - \sum_{s \neq k} \mu_{s,j}\pi_{i,s}$, then:

$$p(\pi_{i,k}|Z_{i,k} = \ell) \propto \exp\left(-\sum_{j=1}^{p} \frac{1}{2\sigma_j}\left(T_{i,k,j} - \pi_{i,k}\mu_{k,j}\right)^2\right) \exp\left(-\frac{1}{2\eta_k}\pi_{i,k}^2\right)$$

$$[\pi_{i,k}|Z_{i,k} = \ell] \sim N\left(\sum_{j=1}^{p} \frac{T_{i,k,j}\mu_{k,j}}{\sigma_j}\left(\sum_{j=1}^{p} \frac{\mu_{k,j}^2}{\sigma_j} + \frac{1}{\eta_k}\right)^{-1}, \left(\sum_{j=1}^{p} \frac{\mu_{k,j}^2}{\sigma_j} + \frac{1}{\eta_k}\right)^{-1}\right)$$

Full conditional of w The full conditional of w_k is defined as $Beta\left(1 + \sum_i 1(Z_{i,k} = 1), 1 + \sum_i 1(Z_{i,k} = 0)\right)$

Full conditional of Z The full conditional of $Z_{i,k}$ is defined as:

$$Z_{i,k} \sim Binomial\left(\frac{w_k N_{[0,1]}(\pi_{i,k}; 0, 0.0001)}{w_k N_{[0,1]}(\pi_{i,k}; 0, 0.0001) + (1 - w_k)N_{[0,1]}(\pi_{i,k}; 0, \gamma_k)}\right)$$

Full conditional of γ The full conditional of γ_w is defined as:

$$\text{Inverse-Gamma}\left(\alpha_\gamma + \sum_i 1(Z_{i,k} = 0)/2, \beta_\gamma + 0.5\sum_{i|Z_{i,k}=0} \pi_{i,k}^2\right)$$

with $\alpha_\gamma = 3$ and $\beta_\gamma = 1$.

3 Synthetic Data

Data Generation The performance of BayesDeBulk was evaluated based on synthetic data generated from a Gaussian model. We considered different simulation scenarios with varying numbers of cell-types, genes and samples; i.e., $K = 10, 20$, $p = 200, 400$, $n = 50, 100$, and variance levels ν and σ. For each synthetic scenario, 30 replicate datasets were generated and the performance of different algorithms was evaluated based on two metrics: Pearson's correlation and mean squared error (MSE) between estimated fractions and true fractions. Figure 1 summarizes how the data was generated. For each cell-type, 20 cell-type-specific markers were randomly sampled from the full list of p genes. Let I_k be the set of cell-type specific markers for the k-th cell. The mean of cell-type specific markers for a particular cell-type k was sampled from a Gaussian distribution with mean randomly drawn from the interval $[1, 3]$ and standard deviation 0.5; while the mean of other markers from a Gaussian distribution centered on zero and standard deviation 0.5. The fraction of different cell-types, i.e., $(\pi_{1,i}, \ldots \pi_{K,i})$, was randomly generated from a Dirichlet distribution with parameter 0.5. Given these parameters, mixed data for the i-th sample was generated as follows:

$$\boldsymbol{Y}_i = \pi_{1,i}\boldsymbol{V}_{1,i} + \ldots \pi_{K,i}\boldsymbol{V}_{K,i} + \boldsymbol{\epsilon}_i$$

with $\boldsymbol{\epsilon}_i \sim N(0, \nu I)$ and $\boldsymbol{V}_{k,i} \sim N(\boldsymbol{\mu}_k, \sigma I)$.

Prior knowledge We implemented different algorithms for different degrees of prior knowledge on cell-type-specific markers where, for each cell-type, (i) 100% of cell-type-specific markers are known a priori and (ii) 50% of cell-type specific markers are known. Specifically, under (ii), 50% of known cell-type-specific markers were randomly drawn from the original set of cell-type specific markers. Cibersort and Epic require as input the gene expression of markers for different cell types (referred to as signature matrix). For a fair comparison, a perturbed version of the original signature matrix was considered as input (Figure 1 B). This signature matrix was generated following two approaches: (i) considering the same set of cell-type specific markers, the signature matrix was regenerated (Figure 1 B); (ii) the mean of 50% of cell-type-specific markers was randomly drawn from the interval $[1, 3]$ while the expression of the remaining 50% was sampled

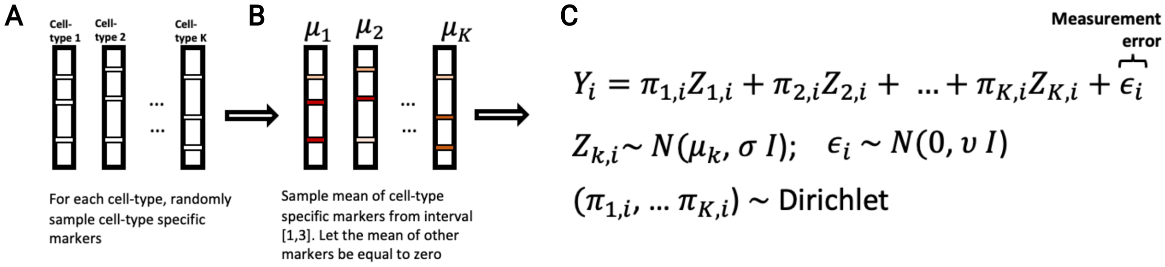
A
Cell-type 1
Cell-type 2
Cell-type K
...
For each cell-type, randomly sample cell-type specific markers
B
μ_1
μ_2
μ_K
...
Sample mean of cell-type specific markers from interval [1,3]. Let the mean of other markers be equal to zero
C
Measurement error
$Y_i = \pi_{1,i}Z_{1,i} + \pi_{2,i}Z_{2,i} + \ldots + \pi_{K,i}Z_{K,i} + \overbrace{\epsilon_i}$
$Z_{k,i} \sim N(\mu_k, \sigma I)$; $\epsilon_i \sim N(0, \upsilon I)$
$(\pi_{1,i}, \ldots \pi_{K,i}) \sim$ Dirichlet

Figure 1: Synthetic data generation.

from a Normal with mean zero and standard deviation 0.5. Basically, after perturbation, 100% of cell-specific markers will be upregulated in a particular component following (i); while only 50% following (ii). Figure 2 shows the scatterplot between the original signature matrix utilized to generate the data and the perturbed signature matrices for both $K = 10$ and $K = 20$ cell-types simulation scenarios. BayesDeBulk was compared with Cibersort [8], Plier [7], xCell [2] and EPIC [11] based on different simulation scenarios with varying numbers of cell-types and markers; i.e., $(K, p, n) = (10, 200, 50)$, $(K, p, n) = (20, 400, 50)$, $(K, p, n) = (10, 200, 100)$ and $(K, p, n) = (20, 400, 100)$, and variance levels ν and σ. For each synthetic scenario, 30 replicate datasets were generated and the performance of different models was evaluated based on two metrics: Pearson's correlation and mean squared error (MSE) between estimated fractions and true fractions.

Algorithms implementation BayesDeBulk: For each synthetic data scenario, BayesDeBulk was estimated considering $10,000$ Marcov Chain Monte Carlo (MCMC) iterations; with the estimated fractions being the mean across iterations after discarding a burn-in of $1,000$. For each sample i, once estimated, cell-type fractions $\{\pi_{i,k}\}_{k=1}^{K}$ were standardized to sum to 1. In addition to cell-type fractions, BayesDeBulk can perform the estimation of the mean parameter for different cell types. The mean parameter $\mu_{j,k}$ was estimated for each gene j and cell k as the mean across 10,000 MCMC iterations after discarding a burn-in of 1,000 iterations. Plier: Plier requires as input a matrix containing cell-type specific markers. Each factor is then modeled as a function of this prior knowledge. The number of estimated factors was fixed to the total number of cell-types. One problem with Plier is that, once estimated, factors might map to multiple cell-types or to none of them. We only considered replicates for which all the factors could be uniquely mapped to a cell-type. The median number of replicates satisfying this requirement across all synthetic scenarios was 23 out of 30 with an interquartile range of $(15, 27)$. xCell xCell was implemented using the Bioconductor package GSVA [4]. Cibersort For each synthetic data, Cibersort was implemented using $P = 100$ permutations and the relative mode. Quantile normalization was not performed.

Results Figure 3 shows the performance of all the models in estimating cell-type fractions for different synthetic data scenarios considering a sample size $n = 100$ under the assumption that 100% of cell-type specific markers is known. As shown, BayesDeBulk resulted in the highest Pearson's correlation and lowest MSE between estimated and true cell fractions for different synthetic data scenarios. Figures 4 and 5 show the performance of different models in estimating cell-type fractions for different synthetic data scenarios considering a sample size $n = 50$, under the assumption that 50% and 100% of cell-type specific markers is known. Again, BayesDeBulk results in the best performance in terms of both Pearson's correlation and mean squared error between true and estimated values.

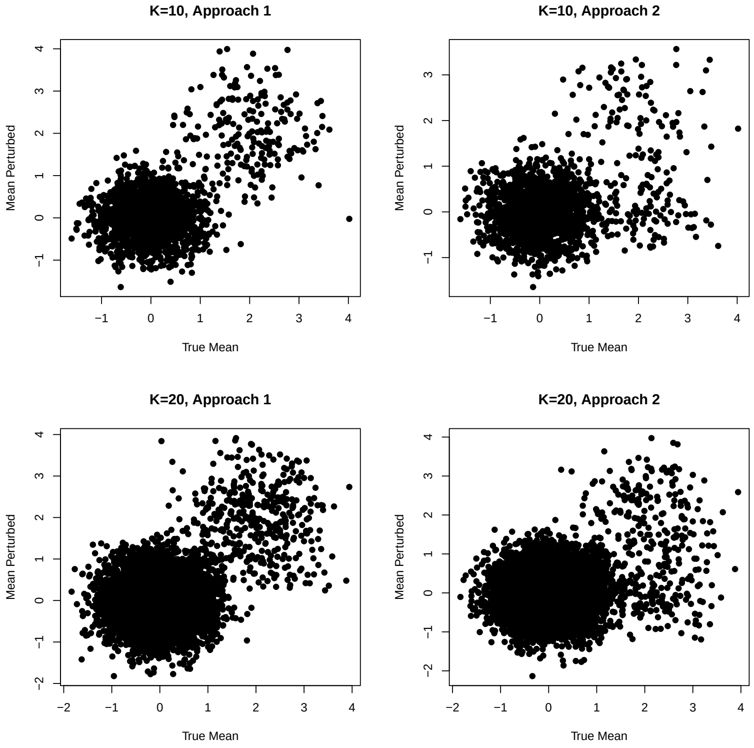

Figure 2: Scatterplot of perturbed signature matrix (y-axis) versus true signature matrix (x-axis) based on 10 cell-types (first row) and 20 cell-types simulation scenarios (second row). Perturbed signatures derived based on approach 1 are shown on the first column; while those derived via approach 2 on the second column.

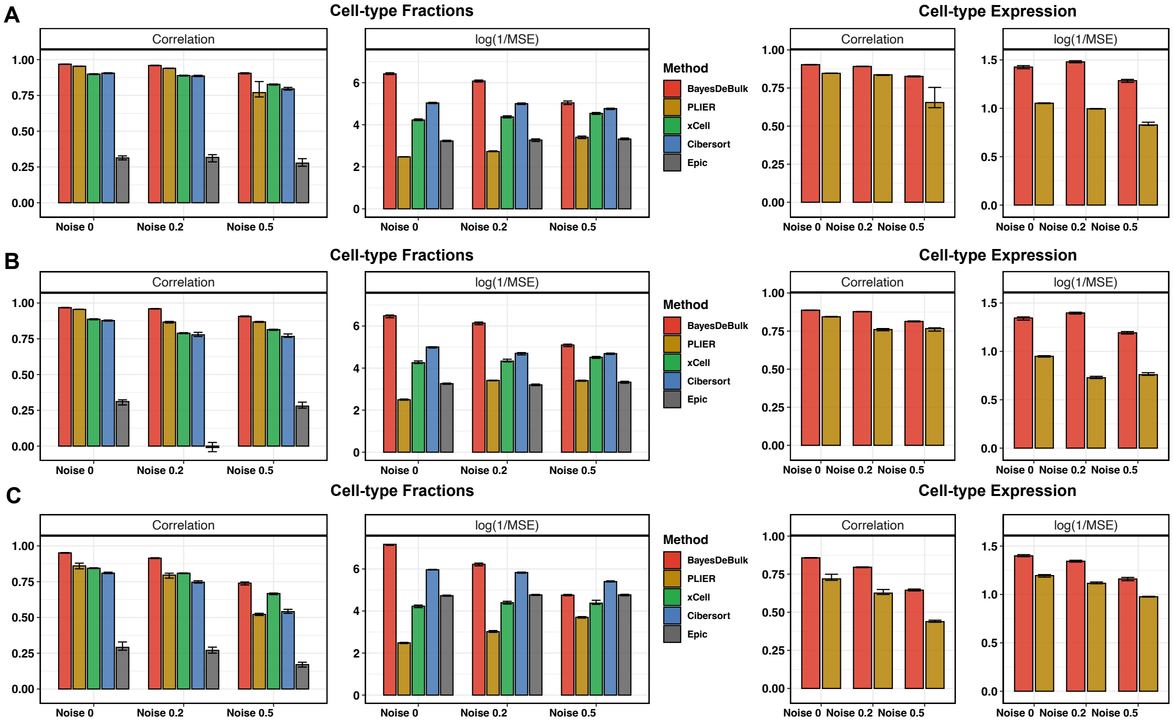

Figure 3: **Synthetic data involving 100 samples under the assumption 100% of cell-type specific markers are known** Pearson's correlation and mean squared error (MSE) between estimated values and truth over 30 replicates for BayesDeBulk (red), Cibersort (blue), xCell (green), Epic (gray) and Plier (gold). Barplots correspond to the median across different replicates while error bars to the interquartile range. For each simulation scenario, we report the correlation and MSE between the estimated cell-type fractions and truth (left-hand panel) for all five algorithms, and between the estimated cell-type expression and the truth (right-hand panel) for BayesDeBulk and Plier. Results are based on data simulated for (A) $K = 10$ and $\sigma = 0.5$; (B) $K = 10$ and $\sigma = 1$; (C) $K = 20$ and $\sigma = 0.5$ for different level of measurement errors ν (noise).

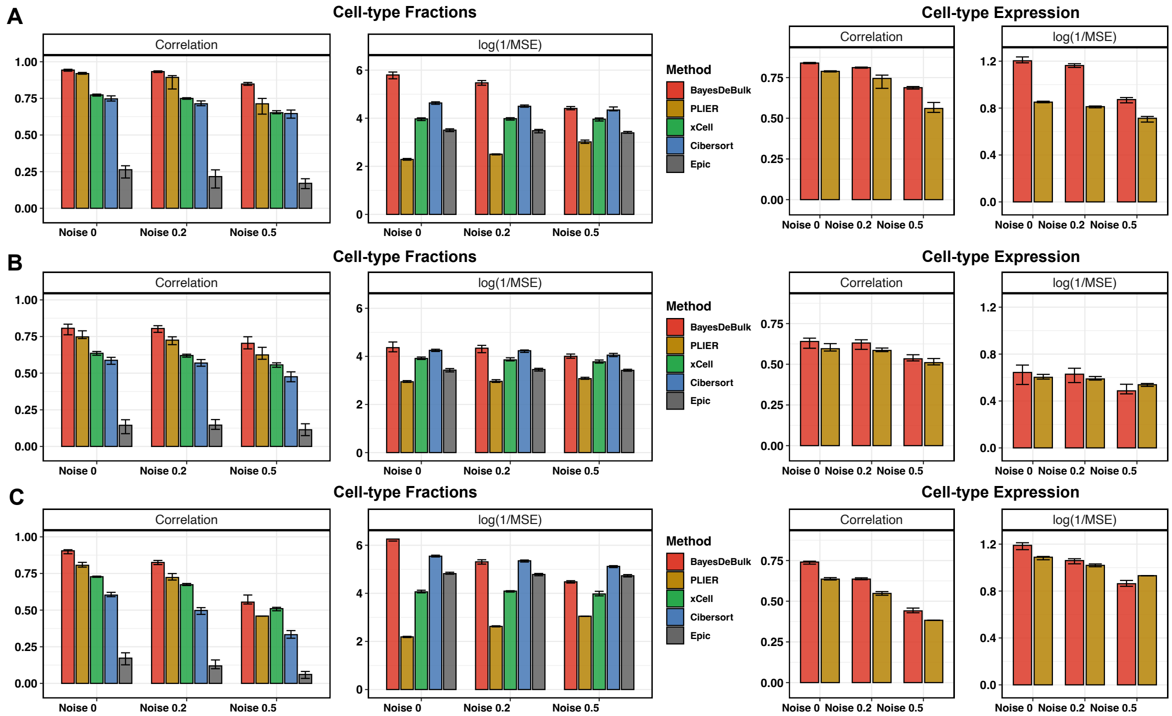

Figure 4: **Synthetic data involving 50 samples under the assumption 50% of cell-type specific markers are known** Pearson's correlation and mean squared error (MSE) between estimated values and truth over 30 replicates for BayesDeBulk (red), Cibersort (blue), xCell (green), Epic (gray) and Plier (gold). Barplots correspond to the median across different replicates while error bars to the interquartile range. For each simulation scenario, we report the correlation and MSE between the estimated cell-type fractions and truth (left-hand panel) for all five algorithms, and between the estimated cell-type expression and the truth (right-hand panel) for BayesDeBulk and Plier. Results are based on data simulated for (A) $K = 10$ and $\sigma = 0.5$; (B) $K = 10$ and $\sigma = 1$; (C) $K = 20$ and $\sigma = 0.5$ for different level of measurement errors ν (noise).

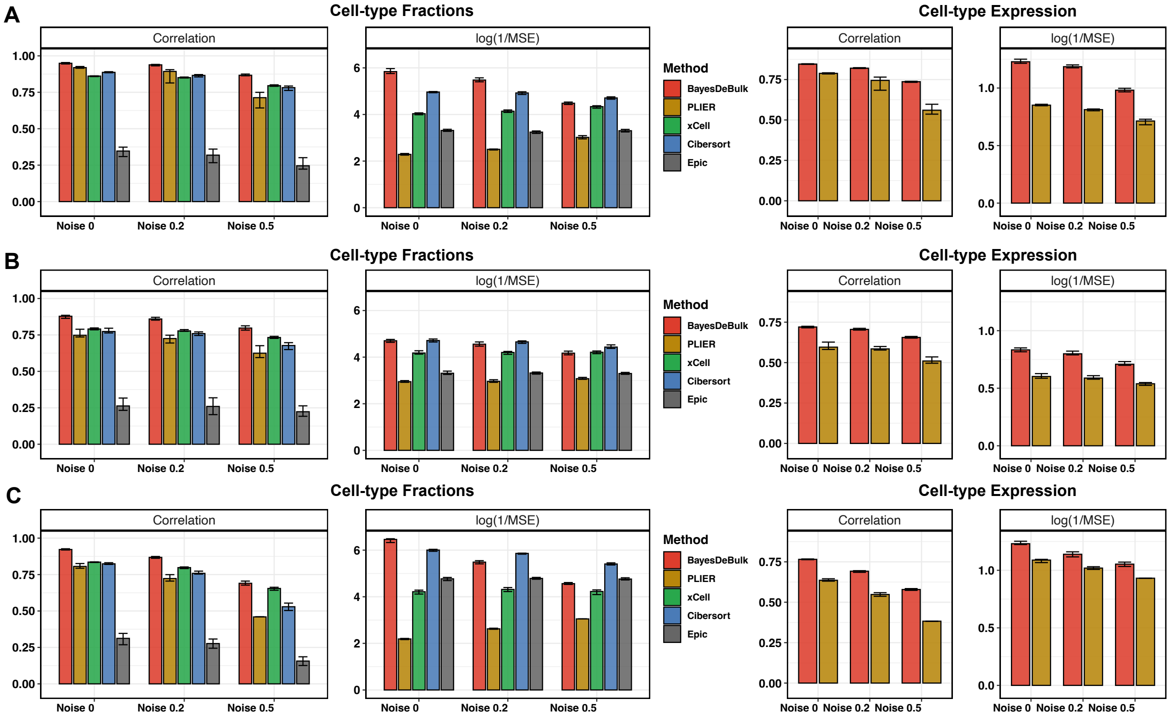

Figure 5: **Synthetic data involving 50 samples under the assumption 100% of cell-type specific markers are known** Pearson's correlation and mean squared error (MSE) between estimated values and truth over 30 replicates for BayesDeBulk (red), Cibersort (blue), xCell (green), Epic (gray) and Plier (gold). Barplots correspond to the median across different replicates while error bars to the interquartile range. For each simulation scenario, we report the correlation and MSE between the estimated cell-type fractions and truth (left-hand panel) for all five algorithms, and between the estimated cell-type expression and the truth (right-hand panel) for BayesDeBulk and Plier. Results are based on data simulated for (A) $K = 10$ and $\sigma = 0.5$; (B) $K = 10$ and $\sigma = 1$; (C) $K = 20$ and $\sigma = 0.5$ for different level of measurement errors ν (noise).

4 Validation based on cytometry immunoprofile

Prior knowledge For the implementation of BayesDeBulk, Epic [11] , Plier [7] and Cibersort [8], the LM22 signature matrix from Cibersort was considered [8]. For MCP-counter [3] and xCell [2] estimation, their default signatures were utilized. Contrary to other algorithms, BayesDeBulk can take as input the list of markers upregulated in one cell-type compared to another cell-type. This is a more flexible strategy than requiring a marker to be higher expressed in one cell-type compared to all other cell-types. For each pair of cells (k, s), we considered a marker ℓ upregulated in the k-th cell type compared to the s-th cell type if $(L_k^\ell > 5 \times L_s^\ell)$ and $L_k^\ell > 1000$ with L_k^ℓ being the value in the LM22 signature matrix of the ℓ marker for the k-th cell type. On the other hand, Plier requires a list of markers expressed in each cell-type. For its implementation, a marker ℓ was considered expressed in the k-th cell-type if $L_k^\ell > 1000$. For the implementation of EPIC and Cibersort, the originial LM22 signature matrix was considered as input.

Estimation For the implementation of BayesDeBulk and Plier, the data was first log transformed and each gene was standardized to z-score (mean zero and standard deviation 1). For BayesDeBulk deconvolution, the LM22 signature matrix values were considered as prior mean $\{\xi_{k,j}\}$ (see section 3.1 in the main manuscript); while $\lambda_{k,j}$ was set to 1. BayesDeBulk model was estimated considering $10,000$ MCMC iterations; with the estimated fractions derived as the mean across iterations after discarding a burn-in of $1,000$. Once estimated, parameters $\{\pi_{i,k}\}_{k=1}^K$ were standardized to sum to 1 for each sample i. For CibersortX based deconvolution, a batch correction step was implemented (B-batch mode). The relative mode was utilized for both Cibersort and CibersortX deconvolutions.

5 Mixture of protein/gene expression from purified cells

Mixture of RNA expression from pure cells Our simulation framework relied on two published datasets. First, we considered data from [6] which contains transcriptomic profile of $K = 6$ immune cell types such as Neutrophil, Natural Killers, B cells, CD4 T cells, CD8 T cells and Monocytes. Let μ_k be the averaged transcriptomic data across multiple replicates for the k-th cell type. For each sample n, weights of different immune cells were randomly sampled from a Dirichlet distribution with parameter 0.5 (i.e., $\pi_{n,1}, \pi_{n,2} ... \pi_{n,K}$). Count data was first log2 transformed and then mixed data was derived as the weighted average of transcriptomic profile of different cell-types as follows $y_n = \sum_{k=1}^K \pi_{n,k} Z_{n,k} + \epsilon_n$ with $Z_{n,k} \sim N(\mu_k, \sigma)$ with $\sigma = sd(\mu)/2$ and $\epsilon_n \sim N(0, \delta)$, and δ being chosen to ensure a 1:1 signal to noise ratio.

Mixture of protein expression from pure cells We considered data from Rieckmann et al [12] including proteomic profile of the same set of immune cells. Considering the same set of weights $\{\pi_{n,k}\}$, mixed proteomic data was generated in a similar fashion as the transcriptomic profile.

Prior knowledge For the implementation of BayesDeBulk, Epic [11] , Plier [7] and Cibersort [8], the LM22 signature matrix from Cibersort was considered. For MCP-counter [3] and xCell [2] estimation, their default signatures were utilized. Contrary to other algorithms, BayesDeBulk can take as input the list of markers upregulated in one cell-type compared to another cell-type. This is a more flexible strategy than requiring a marker to be higher expressed in one cell-type compared to all other cell-types. For each pair of cells (k, s), we considered a marker ℓ upregulated in the k-th cell type compared to the s-th cell type if $(L_k^\ell > 5 \times L_s^\ell)$ and $L_k^\ell > 1000$ with L_k^ℓ being the value in the LM22 signature matrix of the ℓ marker for the k-th cell type. On the other hand, Plier requires a list of markers expressed in each cell-type. For its implementation, a marker ℓ was considered expressed in the k-th cell-type if $L_k^\ell > 1000$. For the implementation of EPIC and Cibersort, the originial LM22 signature matrix was considered as input.

References

[1] Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pp. 1100–1120, 1982.

[2] Dvir Aran, Zicheng Hu, and Atul J Butte. xcell: digitally portraying the tissue cellular heterogeneity landscape. *Genome biology*, 18(1):1–14, 2017.

[3] Etienne Becht, Nicolas A Giraldo, Laetitia Lacroix, Bénédicte Buttard, Nabila Elarouci, Florent Petitprez, Janick Selves, Pierre Laurent-Puig, Catherine Sautès-Fridman, Wolf H Fridman, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome biology*, 17(1):1–20, 2016.

[4] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics*, 14(1):1–15, 2013.

[5] Haiqun Lin and Daniel Zelterman. Modeling survival data: extending the cox model, 2002.

[6] Peter S Linsley, Cate Speake, Elizabeth Whalen, and Damien Chaussabel. Copy number loss of the interferon gene cluster in melanomas is linked to reduced t cell infiltrate and poor patient prognosis. *PloS one*, 9(10):e109760, 2014.

[7] Weiguang Mao, Elena Zaslavsky, Boris M Hartmann, Stuart C Sealfon, and Maria Chikina. Pathway-level information extractor (plier) for gene expression data. *Nature methods*, 16(7):607–610, 2019.

[8] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457, 2015.

[9] Francesca Petralia, Vinayak Rao, and David Dunson. Repulsive mixtures. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[10] Francesca Petralia, Li Wang, Jie Peng, Arthur Yan, Jun Zhu, and Pei Wang. A new method for constructing tumor specific gene co-expression networks based on samples with tumor purity heterogeneity. *Bioinformatics*, 34(13):i528–i536, 2018.

[11] Julien Racle and David Gfeller. Epic: a tool to estimate the proportions of different cell types from bulk gene expression data. In *Bioinformatics for Cancer Immunotherapy*, pp. 233–248. Springer, 2020.

[12] Jan C Rieckmann, Roger Geiger, Daniel Hornburg, Tobias Wolf, Ksenya Kveler, David Jarrossay, Federica Sallusto, Shai S Shen-Orr, Antonio Lanzavecchia, Matthias Mann, et al. Social network architecture of human immune cells unveiled by quantitative proteomics. *Nature immunology*, 18(5):583–593, 2017.