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

Francesca Petralia[1], Anna P. Calinawan[1], Song Feng [2], Sara Gosline[2],

Pietro Pugliese[3], Michele Ceccarelli[4], Pei Wang[1]

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

Abstract

Characterizing the tumor microenvironment is crucial in order to improve responsiveness to immunotherapy and develop new therapeutic strategies. The fraction of different cell-types in the tumor microenvironment can be estimated based on transcriptomic profiling of bulk tumor data via deconvolution algorithms. One class of such algorithms, known as reference-based, rely on a reference signature containing gene expression data for various cell-types. The limitation of these methods is that such a signature is derived from the gene expression of pure cell-types, which might not be consistent with the transcriptomic profiling in solid tumors. On the other hand, reference-free methods usually require only a set of cell-specific markers to perform deconvolution; however, once the different components have been estimated from the data, their labeling can be problematic. To overcome these limitations, we propose BayesDeBulk - a new reference-free Bayesian method for bulk deconvolution based on gene expression data. Given a list of markers expressed in each cell-type (cell-specific markers), a repulsive prior is placed on the mean of gene expression in different cell-types to ensure that cell-specific markers are upregulated in a particular component. Contrary to existing reference-free methods, the labeling of different components is decided a priori through a repulsive prior. Furthermore, the advantage over reference-based algorithms is that the cell fractions as well as the gene expression of different cells are estimated from the data, simultaneously. Given its flexibility, BayesDeBulk can be utilized to perform bulk deconvolution beyond transcriptomic data, based on other data types such as proteomic profiles or the integration of both transcriptomic and proteomic profiles.

Contents

1 Introduction

Solid tumors are composed of a variety of cell-types including immune and stromal cells. Quantifying the proportion of different cell-types in the tumor microenvironment is crucial in order to capture patient heterogeneity and develop better therapeutic targets for precision medicine. In the last decade, different algorithms have been proposed for the estimation of the tumor microenvironment from bulk data. Some algorithms, known as reference-based, require gene expression of purified cells as prior information [2, 9, 18]. However, problems might arise when the gene expression of different cell-types in solid tumors are not consistent with this prior knowledge. In addition, this prior information might not be appropriate when performing the deconvolution based on other data types such as proteomic profiles.

Some algorithms have tried to overcome this lack of flexibility by proposing a semi-reference approach [3, 11, 16]. Recently, Tai et al, [16] proposed a semi-reference Bayesian method which jointly models gene expression from purified cells and that from bulk data through a hierarchical model. This model is more flexible than reference-based methods since gene expression in different cell-types is inferred from the data; however, it relies on the assumption that the mean expression of a particular cell in the reference data and in bulk data are the same. Aran et al (2017) [1] proposed a flexible tool for bulk deconvolution based on transcriptomic data, which requires only a list of markers expressed in each cell-type as prior information. However, this algorithm does not provide an estimate of the gene expression of different cell types from the data. These estimates are particularly useful in order to perform differential expression analyses between tumor and adjacent normal tissues while accounting for immune and stromal infiltration.

To this end, there are many reference-free algorithms which can infer both cell-type proportions and gene expression in different cell-types [5, 6, 8, 14]. These algorithms represent a more flexible alternative to reference-based algorithms; however, once estimated, the interpretation and the labeling of different components might be problematic. Recently, Tang et al [17] proposed an algorithm based on non-negative matrix factorization. This method recovers the identifiability and the labeling of different components using a penalized regression, in which markers expected to be less expressed in a particular cell type shrink towards zero. For this purpose, for each cell-type, markers are divided into three categories: not expressed, expressed and highly expressed. However, marker stratification into such categories might not be easy to achieve in practice.

To overcome these limitations, we propose BayesDeBulk - a new flexible Bayesian method for bulk deconvolution. Bayesian inference is very appealing in this framework since prior information for different cell-types can be flexibly incorporated through the prior. Given a list of markers expressed in a particular cell-type (cell-specific markers), a repulsive prior is placed on the mean of gene expression in different cell-types to ensure that cell-specific markers are upregulated in a particular component. Repulsive classes of priors have been introduced by Petralia et al [10]; and recently extended to different applications [12, 13, 19, 20]. Contrary to existing reference-free methods, the labeling of different components is specified a priori through a repulsive prior. The cell fraction parameter is instead modeled through a spike-and-slab prior [4] in order to induce sparsity and identify cells which are not present in the tumor tissue. Contrary to reference-based algorithms, our framework estimates different cell-type fractions and the mean of gene expression in different cell-types from the data, simultaneously. Given its flexibility, BayesDeBulk can be used to perform the deconvolution based on other data types such as methylation data and proteomic profiles or the integration of multi-omic data. The performance of our model is evaluated using extensive synthetic data and real data examples.

2 Method

2.1 Bulk Deconvolution

Since the expression of bulk tumor data is the average across different cells in the tumor microenvironment, the expression of gene j for patient i, i.e., $y_{i,j}$, can be modeled as a Gaussian distribution with mean parameter being the weighted average between the expression of gene j in different cell-types. Mathematically, $y_{i,j}$ is modeled as

$$y_{i,j} \sim N(\theta_{i,j}, \sigma_j) \ \ , \ \ \theta_{i,j} = \sum_{k=1}^{K} \pi_{i,k}\mu_{k,j}$$

with K being the total number of cell-types, $\pi_{i,k}$ being the fraction of the k-th cell-type for sample i, $\mu_{k,j}$ being the expression of gene j for the k-th cell-type and σ_j the variance of the j-th gene. Reference-based models would consider $\mu_{k,j}$ as fixed with measurements derived from existing pure cell transcriptomic data [2, 9, 18]; while reference-free models would estimate mean parameters $\{\mu_{k,j}\}$ from the bulk data. A Bayesian model would specify prior information for all parameters in the model; with conjugate priors being Gaussian distributions for $\{\pi_{i,k}\}_{k=1}^{K}$ and $\{\mu_{k,j}\}$ and inverse-gamma distributions for $\{\sigma_j\}_{j=1}^{p}$. However, this model would not be identifiable without further constraints on the parameter space. To overcome this problem we propose a Bayesian model where identifiability is recovered via a repulsive prior specified on the mean parameters [10].

2.2 Bayesian model based on repulsive prior

Let us assume that for each k-th cell-type, there is a set I_k of genes whose expression is upregulated in the k-th cell-type compared to all others. We will use a flexible repulsive prior [10] in order to ensure that genes in set I_k will have a ”larger” mean in the k-th cell-type compared to other cell-types. Let $\boldsymbol{\mu}_k$ be a p dimensional vector containing the gene expression of p genes in the kth cell-type. Then, $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ is jointly modeled through the following multivariate prior:

$$p(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K) = \prod_{k=1}^{K} \left[\prod_{j=1}^{p} N(\mu_{k,j}; 0, 1) \right] h(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$$

with $h(-)$ being a repulsive function defined as

$$h(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K) = \min_{\forall s} \min_{\forall j \in I_s; \forall k \neq s} \exp(-\tau(|\mu_{s,j} - \mu_{k,j}|)^{-\eta}) 1(\mu_{s,j} > \mu_{k,j})$$

with $\tau > 0$ and $\eta > 0$. This function is an extension of the repulsive function introduced by Petralia et al [10], and it approaches zero as the distance between mean parameters goes to zero and the upregulation of genes belonging to set I_s in the s-th cell is not satisfied. According to this function, genes contained in set I_k will have a mean value greater in component k-th compared to all other components. It is important to note that only genes contained in set $I = \cup_{k=1}^{K} I_k$ will be assigned a repulsive prior; other genes will have a standard normal prior. This is sufficient to recover identifiability of the model and will reduce substantially the computational burden. Prior knowledge on markers upregulated in each cell-type can be leveraged from existing databases and single cell RNA data. Instead of requiring a set of markers to be upregulated in one cell-type compared to all other cell-types; the user might specify this requirement for each pair of cells. For instance, assume that $I_{s>k}$ is the set of genes upregulated in the s-th cell-type compared to the k-th cell-type. In this case, the repulsive prior can be easily modified to incorporate this information in the following way:

$$h(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K) = \min_{\forall s} \min_{\forall j \in I_{s>k}; \forall k \neq s} \exp(-\tau(|\mu_{s,j} - \mu_{k,j}|)^{-\eta}) 1(\mu_{s,j} > \mu_{k,j})$$

To facilitate computation, we will not require $\{\pi_{i,k}\}_{k=1}^{K}$ to sum to 1. However, we will require these parameter to be defined on the unit interval $[0,1]$. As prior specification, we will use a spike-and-slab prior [4] defined on the unit interval, i.e., $\pi_{i,k} \sim w_k N_{[0,1]}(0, 0.0001) + (1 - w_k) N_{[0,1]}(0, \gamma_k)$ with $w_k \sim Beta(1,1)$ and $\gamma_k \sim$ Inverse-Gamma(a_γ, b_γ). The spike component concentrates its mass at values close to zero, shrinking small effects to zero, and therefore inducing sparsity in cell fractions estimates. The percentage of zero values (i.e., w_k) will vary across different cell-types. We expect that some cell-types will be more abundant (i.e., different from zeros) than others in a particular tissue. For instance, T cells will be more likely present in kidney or lung tissues rather than brain tissues. For the variance components $\{\sigma_j\}_{j=1}^{p}$, standard inverse-gamma priors will be utilized. Figure 1 provides a summary of the proposed model.

This algorithm can be easily utilized to perform a multi-omic based deconvolution. In this case, each data type can be modeled via a BayesDeBulk model, with different data-specific models sharing the same set of cell fraction parameters. Let $y_{i,j}$ and $z_{i,j}$ be the RNA and protein expression of gene j for sample i. A multi-omic framework would model $y_{i,j}$ and $z_{i,j}$ as

$$y_{i,j} \sim N(\theta_{i,j}, \sigma_j) \quad , \quad \theta_{i,j} = \sum_{k=1}^{K} \pi_{i,k} \mu_{k,j}$$

$$z_{i,j} \sim N(\beta_{i,j}, \iota_j) \quad , \quad \beta_{i,j} = \sum_{k=1}^{K} \pi_{i,k} \delta_{k,j}$$

with K being the total number of cell-types, $\pi_{i,k}$ being the fraction of the k-th cell-type for sample i, $\mu_{k,j}$ being the expression of gene j for the k-th cell-type, $\delta_{k,j}$ being the expression of protein j for the k-th cell-type. It is important to notice that the two models share the same set of cell fractions, i.e. $\{\pi_{i,k}\}$. In this case, the same repulsive prior would be placed on the mean parameters of both models, i.e., $\{\mu_{k,j}\}$ and $\{\delta_{k,j}\}$.

2.3 Full conditionals and posterior computation

Following Petralia et al [10], a latent variable ρ will be introduced to facilitate the sampling from the repulsive prior. This latent variable will be jointly modeled with $\boldsymbol{\mu}$ through the following multivariate density:

$$p(\boldsymbol{\mu}, \rho) = \left[\prod_{\forall s; \forall j} \mathrm{N}(\mu_{s,j}; 0, 1) \right] 1(h(\boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_K) > \rho)$$

A set of additional latent variables $\{Z_{i,k}\}$ will be introduced in order to facilitate the sampling from the spike-and-slab prior placed on $\{\pi_{i,k}\}$. In particular, $Z_{i,k}$ will be equal to 1 if $\pi_{i,k}$ will be sampled from the "spike" component, i.e., $\pi_{i,k} \sim N_{[0,1]}(0, 0.0001)$; while equal to 0 if $\pi_{i,k}$ will be sampled from the "slab" component, i.e., $\pi_{i,k} \sim N_{[0,1]}(0, \gamma_k)$. Let $1(A)$ be an indicator function equal to 1 if A is satisfied and 0 otherwise. The Gibbs sampler can be summarized in the following steps.

Step 1 Sample mean parameter $\mu_{k,j}$ from a truncated normal distribution:

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

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 1 of the supplementary material.

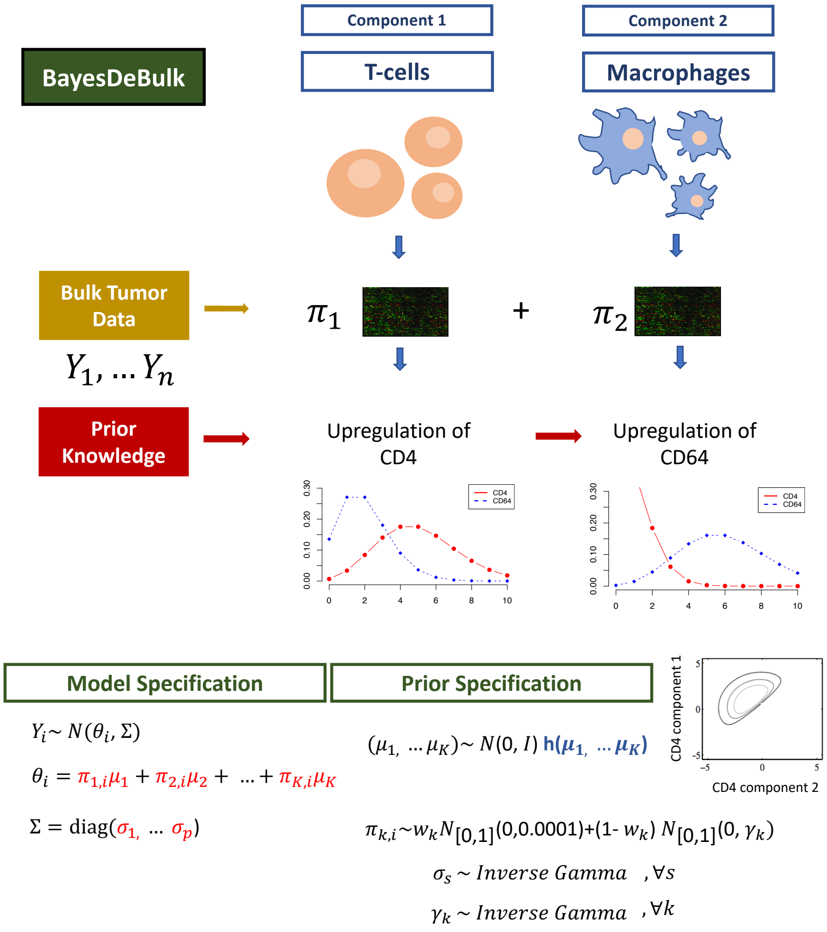
BayesDeBulk
Component 1
T-cells
Component 2
Macrophages
Bulk Tumor Data
$Y_1, \dots Y_n$
π_1
+
π_2
Prior Knowledge
Upregulation of CD4
Upregulation of CD64
CD4
CD64
0.00
0.10
0.20
0.30
0
2
4
6
8
10
Model Specification
Prior Specification
$Y_i \sim N(\theta_i, \Sigma)$
$\theta_i = \pi_{1,i}\mu_1 + \pi_{2,i}\mu_2 + \dots + \pi_{K,i}\mu_K$
$\Sigma = \mathrm{diag}(\sigma_{1,} \dots \sigma_p)$
$(\mu_{1,} \dots \mu_K) \sim N(0, I)\ \mathbf{h}(\mu_{1,} \dots \mu_K)$
$\pi_{k,i} \sim w_k N_{[0,1]}(0,0.0001)+(1- w_k)\ N_{[0,1]}(0, \gamma_k)$
$\sigma_s \sim Inverse\ Gamma\ \ , \forall s$
$\gamma_k \sim Inverse\ Gamma\ \ , \forall k$
CD4 component 1
CD4 component 2
-5
0
5

Figure 1: Algorithm Schematic

Step 2 Sample $Z_{i,k}$ from

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

Step 3 Sample $\pi_{i,k}$ from a truncated univariate normal defined as:

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

with $T_{i,k,j}$ being defined as $y_{i,j} - \sum_{s\neq k}\mu_{s,j}\pi_{i,s}$ and $\eta_k = \gamma_k$ if $\ell = 0$ and $\eta_k = 0.0001$ if $\ell = 1$.

Step 4 Sample w_k from $Beta\left(1+\sum_i 1(Z_{i,k}=1), 1+\sum_i 1(Z_{i,k}=0)\right)$
Step 5 Sample γ_k from:

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

Step 6 Sample σ_j from:

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

Step 7 Sample ρ from a uniform distribution

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

Detailed information on how full-conditionals were derived is contained in section 1 of supplementary material.

3 Synthetic Data

3.1 Data Generation

The performance of BayesDeBulk in estimating cell-type fractions and the gene expression in different cells was evaluated based on extensive synthetic data. Let p be the total number of genes, n the total number of samples and K the number of cell-types. Let I_k be the set containing 20 cell-specific markers for the k-th cell; which were randomly sampled from the full list of genes. The mean of cell-specific markers for a particular cell k, i.e., $\mu_{k,j}$ with $j \in I_k$, was drawn from a Gaussian distribution with mean uniformly sampled from the range $[1, 3]$ and standard deviation 0.5; while the mean of other markers, i.e., $\mu_{k,j}$ with $j \notin I_k$, 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 1. 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 $\boldsymbol{V}_{k,i} \sim N(\boldsymbol{\mu}_k, \sigma I)$.

3.2 Results

BayesDeBulk was compared with Cibersort [9] based on different simulation scenarios with varying numbers of cells and genes; i.e., $(K, p, n) = (10, 200, 100)$ and $(K, p, n) = (20, 400, 100)$, and variance levels ν and σ. For each synthetic scenario, 10 replicate datasets were generated and the performance of the two models was evaluated based on two metrics: Pearson's correlation and mean squared error (MSE) between estimated fractions and true fractions. For each replicate, BayesDeBulk was estimated considering 10000 Marcov Chain Monte Carlo (MCMC) iterations; with the estimated fractions being the mean across iterations after discarding a burn-in of 1000. BayesDeBulk was implemented (i) assuming that all cell-specific markers are known a priori (BayesDeBulk 100) and (ii) only 50% of cell-specific markers are known a priori (BayesDeBulk 50). This second scenario is more representative of real world applications, where only a proportion of cell-specific markers is usually known. Contrary to BayesDeBulk, Cibersort requires as input a signature matrix containing the mean of different markers for different cell-types. In order to make a fair comparison, a perturbed version of the original signature matrix was considered as input in Cibersort based inference. Specifically, the original signature matrix was perturbed following two approaches: (i) preserving the upregulation of key cell-specific markers (Cibersort 100), and (ii) preserving only 50% of markers upregulation. The scatterplot between true and perturbed signature matrices can be found in the supplementary material (section 2.1, Supplementary Figure 1).

As shown in Figure 2, BayesDeBulk resulted in a higher Pearson's correlation for different synthetic data scenarios. In particular, a median correlation above 0.90 was observed for BayesDeBulk for all simulation scenarios involving $K = 10$ components; while Cibersort resulted in a median correlation lower than 0.70 for higher noise levels. As expected, the performance of both models decreased as more components were incorporated into the model. Overall, we observe that Cibersort is more sensitive to the prior knowledge incorporated in the model; in fact its performance substantially decreases when only 50% of the markers are known a priori. This is due to the lack of flexibility of Cibersort, which requires as input a signature matrix containing the mean levels of different markers for different cells. Indeed, the advantage of our proposed Bayesian framework is the estimation of the expression of different markers for different cell-types. Section 2.2 of supplementary material shows the performance of BayesDeBulk in estimating the mean of gene expression for different components. As expected, higher noise levels result in lower performance in terms of both correlation and MSE (Supplementary Figures 2, 3). The median Pearson's correlation between estimated and true values across replicates was above 0.80 for the simulations involving 10 cell types; including when only 50% of cell-specific markers are known a priori. Although the median correlation decreases substantially when the number of components increases to $K = 20$, it remains above 0.50 for different simulation scenarios.

4 Real data examples

4.1 Multi-omic based deconvolution

Our simulation framework relied on two published datasets. First, we considered data from [7] which contains transcriptomic profiling 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 profiling of different cell-types as follows $y_n = \sum_{k=1}^{K} \pi_{n,k} Z_{n,k} + \epsilon_n$ with $Z_{n,k} \sim N(\mu_k, \sigma)$ and $\epsilon_n \sim N(0, \delta)$. In particular, δ was chosen to ensure a 1:1 signal to noise ratio. Then, we considered data from Rieckmann et al [15] including proteomic profiling 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 profiling. BayesDeBulk was compared with Cibersort [9] in estimating immune cell fractions. For this comparison, different simulation scenarios with varying number of samples; i.e.,

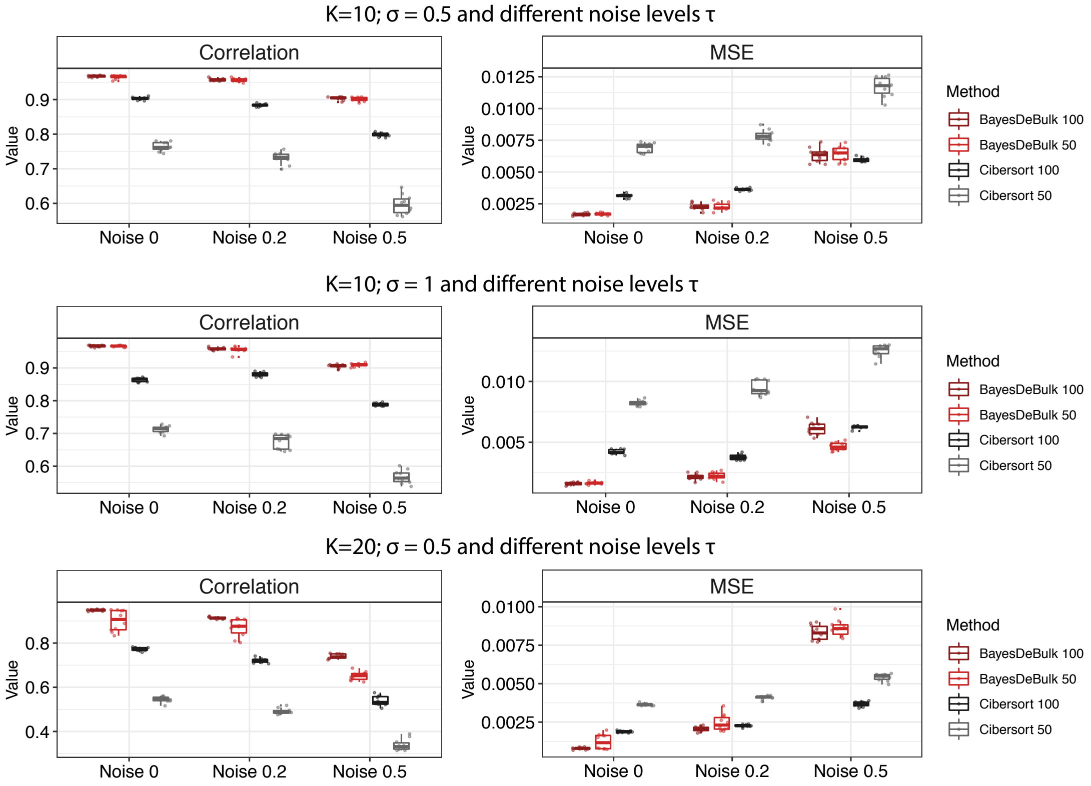

Figure 2: Boxplot of Pearson's correlation and MSE between estimated cell fractions and truth over 10 replicates. Results based on data simulated for (A) $K = 10$ and $\sigma = 0.5$; (B) $K = 10$ and $\sigma = 1$; (C) $K = 20$ and $\sigma = 0.5$ for different level of measurement errors ν. Results based on BayesDeBulk 100, BayesDeBulk 50, Cibersort 100 and Cibersort 50 are shown.

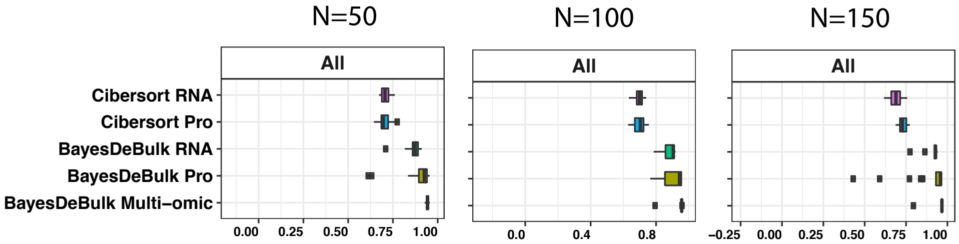

Figure 3: Boxplot of Pearson's correlation between true and estimated cell fractions across 20 replicates for BayesDeBulk and Cibersort models based on different simulation scenarios involving $N = 50$, $N = 100$ and $N = 150$ samples. BayesDeBulk and Cibersort were implemented based on proteomic (Pro) and transcriptomic (RNA) data. For BayesDeBulk, a multi-omic based deconvolution was also performed (Multi-omic).

$(N = 50, N = 100, N = 150)$, were considered. Specifically, for each synthetic scenario, 20 replicate datasets were generated and the performance of the two models was evaluated based on the Pearson's correlation between estimated and true fractions. For each replicate, 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 $1,000$. BayesDeBulk and Cibersort were implemented based on proteomic and transcriptomic data. For BayesDeBulk, a multi-omic based deconvolution was also performed.

Figure 3 shows the boxplot of Pearson's correlation over 20 replicates between the true and estimated fractions for BayesDeBulk and Cibersort based on different data types and simulation scenarios. The performance for each cell-type can be found in Supplementary Figure 4. As shown, BayesDeBulk results in higher correlation between true and estimated cell-fractions than Cibersort. In particular, the multi-omic based deconvolution can outperform single-omic deconvolutions revealing the advantage of a multi-omic based learning.

4.2 Validation based on flow cytometry

In this section, the performance of BayesDeBulk is compared with Cibersort [9] and xCell [1] based on transcriptomic data from peripheral blood mononuclear cells from 20 adults who received influenza immunization [9]. For inference, BayesDeBulk considered the same set of cell-types used in Cibersort; however, both signatures from Cibersort and xCell were considered as prior information. Detailed information on how cell-type specific markers were identified based on both signatures can be found in Section 3 of supplementary material. BayesDeBulk model was estimated considering 3000 MCMC iterations; with the estimated fractions derived as the mean across iterations after discarding a burn-in of 1000. Figure 4 shows the Pearson's correlation between flow-cytometry estimates and estimates derived via different algorithms. As illustrated, BayesDeBulk outperformed both Cibersort and xCell in the estimation of gamma delta T-cells and monocytes. In addition, BayesDeBulk performed better than xCell in the estimation of NK cells and CD8 T cells. For 5 out of 7 cells, BayesDeBulk resulted in a correlation higher than 0.5; compared to 6 out of 7 for Cibersort and 4 out of 7 for xCell. xCell resulted in estimates equal to zero for gamma delta T cells and NK cells.

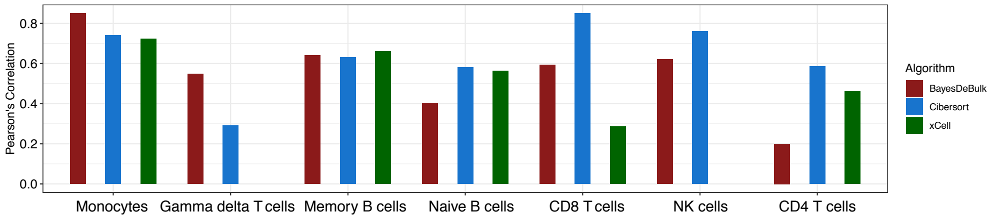

Figure 4: Pearson's correlation between estimated cell fractions and estimates based on flow cytometry data for BayesDeBulk (red), Cibersort (blue) and xCell (green).

5 Conclusion

We introduce BayesDeBulk, a new Bayesian method for the deconvolution of bulk tumor data. BayesDeBulk allows the simultaneous estimation of both cell fractions and gene expression for different cell-types. To perform the deconvolution, BayesDeBulk requires a set of genes expressed in each cell-type, which can be obtained from existing transcriptomic profiles of pure cells. Bulk RNA data is modeled via a Gaussian distribution with mean being the weighted average of expression in different cell-types. Given a list of markers expressed in a particular cell-type, a repulsive prior is placed on the mean of gene expression in different cell-types to ensure that cell-specific markers are upregulated in a particular component. This prior specification facilitates the identification and the labeling of the components contained in the mean parameter; which is a common problem of reference-free methods.

Contrary to reference-based methods, our framework estimates different cell-type fractions and the mean of gene expression in different cell-types from the data, simultaneously. Reference-based algorithms often rely on the assumption that the transcriptomic profiling of different immune/stromal cells in solid tumor is similar to that of the reference data derived from pure cells. Violation of this assumption might lead to poor performance in the estimation of cell fractions. On the other hand, BayesDeBulk does not need to rely on such an assumption since it estimates the transcriptomic profiling of different cells directly from the data.

In addition, the estimation of transcriptomic profiling for different cells is very important in order to perform differential expression analyses between adjacent normal and tumor tissues while accounting for tumor purity. For example, one problem that researchers encounter when performing differential expression analyses between tumor and adjacent normal tissues is that some immune genes might be detected as differentially expressed between tumor and adjacent normal tissues driven by the higher immune infiltration in tumor. BayesDeBulk can be used to estimate the transcriptomic profile of tumor cells by adding an extra component. Then, the estimated profiling of tumor cells might be used in order to identify genes differentially expressed between specifically tumor-cells and adjacent normal tissues.

Given its flexibility, BayesDeBulk can be utilized to characterize the tumor microenvironment based on other data types such as methylation or proteomic profiling. In addition, the algorithm can be easily utilized for a multi-omic based deconvolution. In this case, each data type can be modeled via a BayesDeBulk model, with different data-specific models sharing the same set of cell fraction parameters. This multi-omic framework would allow the estimation of cell fractions based on multi-omic data as well as multi-omic measurements of different markers across different cell-types.

References

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

[2] Binbin Chen, Michael S Khodadoust, Chih Long Liu, Aaron M Newman, and Ash A Alizadeh. Profiling tumor infiltrating immune cells with cibersort. In *Cancer systems biology*, pages 243–259. Springer, 2018.

[3] Li Dong, Avinash Kollipara, Toni Darville, Fei Zou, and Xiaojing Zheng. Semi-cam: a semi-supervised deconvolution method for bulk transcriptomic data with partial marker gene information. *Scientific reports*, 10(1):1–12, 2020.

[4] Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[5] Eugene Andres Houseman, John Molitor, and Carmen J Marsit. Reference-free cell mixture adjustments in analysis of dna methylation data. *Bioinformatics*, 30(10):1431–1439, 2014.

[6] Ziyi Li and Hao Wu. Toast: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome biology*, 20(1):1–17, 2019.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Yufang Qin, Weiwei Zhang, Xiaoqiang Sun, Siwei Nan, Nana Wei, Hua-Jun Wu, and Xiaoqi Zheng. Deconvolution of heterogeneous tumor samples using partial reference signals. *PLOS Computational Biology*, 16(11):e1008452, 2020.

[12] José J Quinlan, Garritt L Page, and Fernando A Quintana. Density regression using repulsive distributions. *Journal of Statistical Computation and Simulation*, 88(15):2931–2947, 2018.

[13] José J Quinlan, Fernando A Quintana, and Garritt L Page. Parsimonious hierarchical modeling using repulsive distributions. *arXiv preprint arXiv:1701.04457*, 2017.

[14] Dirk Repsilber, Sabine Kern, Anna Telaar, Gerhard Walzl, Gillian F Black, Joachim Selbig, Shreemanta K Parida, Stefan HE Kaufmann, and Marc Jacobsen. Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach. *BMC bioinformatics*, 11(1):1–15, 2010.

[15] 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.

[16] An-Shun Tai, George C Tseng, and Wen-Ping Hsieh. Bayice: A bayesian hierarchical model for semireference-based deconvolution of bulk transcriptomic data. *The Annals of Applied Statistics*, 15(1):391–411, 2021.

[17] Daiwei Tang, Seyoung Park, and Hongyu Zhao. Nitumid: nonnegative matrix factorization-based immune-tumor microenvironment deconvolution. *Bioinformatics*, 36(5):1344–1350, 2020.

[18] Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, 10(1):1–9, 2019.

[19] Fangzheng Xie and Yanxun Xu. Bayesian repulsive gaussian mixture model. *Journal of the American Statistical Association*, 115(529):187–203, 2020.

[20] Yanxun Xu, Peter Müller, and Donatello Telesca. Bayesian inference for latent biologic structure with determinantal point processes (dpp). *Biometrics*, 72(3):955–964, 2016.