bioRxiv preprint doi: https://doi.org/10.1101/2021.06.25.449763; this version posted August 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

**Abstract**

**Motivation:** To improve cancer immunotherapy response, one crucial step is to study the immune/stromal cell composition in the tumor microenvironment. The fraction of different cell types in the tumor microenvironment can be estimated via deconvolution algorithms from bulk transcriptomic profiles. One class of such algorithms, known as reference-based, requires as input a reference signature matrix containing the gene expression measurements of different cell types. The limitation of these algorithms is that problems might arise when the transcriptomic profiles of different cell types in solid tumors deviate from the reference, leading to poor estimation performance. A more flexible alternative is given by reference-free methods which can perform the simultaneous estimation of cell-type fractions and cell-type gene expression from the data. However, most of these algorithms rely on factor modeling which unfortunately suffers from interpretability issues, as the labeling of different factors into cell types is often problematic.

**Results:** To overcome these limitations, we propose BayesDeBulk, a novel reference-free Bayesian model which flexibly leverages existing information of known cell-type specific markers and performs the simultaneous estimation of cell-type fractions and cell-type gene expression. Specifically, BayesDebulk imposes a novel Repulsive prior distribution on the mean of cell-type specific markers to ensure the upregulation of cell-type specific markers in a particular component. Using this prior specification, each component of the mixture model is identifiable and automatically assigned to a particular cell type, overcoming the identifiability issues affecting reference-free methods. This flexible framework enables BayesDeBulk to perform the deconvolution by integrating proteomic and transcriptomic data measured for the same set of samples. Improved performance of BayesDeBulk over state-of-the-art deconvolution algorithms such as Cibersort and xCell is shown on different synthetic and real data examples.

**Availability:** Software available at http://www.bayesdebulk.com/

**Contact:** For any information, please contact francesca.petralia@mssm.edu

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.25.449763; this version posted August 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# 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. 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 [4, 18, 34]. However, problems might arise when the gene expression of different cell types in certain organs or tumors does not align with those in purified cells. In addition, this prior information might not be appropriate when performing the deconvolution based on other data types such as proteomic profiles. In recent cancer studies [5, 8, 21, 29], it is becoming common to collect multi-omic data including gene expression and global proteomic data for the same set of biological samples. In this context, it is crucial to fully leverage all available data in order to better estimate the immune and stromal composition in the tumor microenvironment.

To overcome this lack of flexibility, efforts have been made to use semi-reference approaches [6, 22, 31]. Recently, Tai et al, [31] proposed a Bayesian algorithm 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 still relies on the assumption that the mean expression of a particular cell type in the reference data is similar to that in the bulk data.

Aran et al (2017) [1] proposed xCell, a flexible tool for bulk deconvolution based on transcriptomic data. xCell considers as prior information a list of markers expressed in each cell type, and returns an enrichment score reflecting the amount of different cell types in the tumor microenvironment. Although more flexible than other deconvolution algorithms, xCell does not estimate 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 composition.

There are a few other reference-free algorithms that are able to infer both cell-type proportions and marker expression in different cell types [10, 11, 15, 26]. These algorithms utilize factor analysis or similar strategies to avoid the usage of reference signatures. While more flexible, the labeling and interpretation of different components (factors) might be problematic. Recently, Tang et al [32] proposed an algorithm based on non-negative matrix factorization. This method recovers the identifiability and the labeling of different components using a penalized regression, where the gene expression of markers expected to be less expressed in a particular cell type are shrunk towards zero. This algorithm requires that, for each cell type, markers are divided into three categories: not expressed, expressed and highly expressed. A possible drawback of such algorithm is that 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. Given a list of markers expressed in a particular cell type (cell-type specific markers), BayesDebulk imposes a repulsive prior distribution on the mean of marker expression in different cell types to ensure that cell-type specific markers are upregulated in the corresponding cell-

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.25.449763; this version posted August 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

type component. Repulsive class of priors have been introduced by Petralia et al [20]; and were recently extended to different applications [23, 24, 35, 36]. Contrary to existing reference-free methods, the labeling of different components is specified a priori through a repulsive prior. The cell-type fraction parameter is instead modeled through a spike-and-slab prior [7] in order to induce sparsity and identify cell types which are not present in the tumor microenvironment. 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. Furthermore, the labeling of different components is specified a priori, which helps to avoid the ambiguity issue encountered by factor models. Finally, the flexible framework of BayesDeBulk naturally supports the joint modeling of multi-omic data, such as proteomic and transcriptomic data, measured for the same set of samples. The performance of BayesDeBulk is compared to existing deconvolution methods such as Cibersort [18], CibersortX [19], Epic [25], Plier [15], xCell [1] and MCP-counter [2] using extensive synthetic data and real data experiments.

# 2 Background

Gene expression data from bulk experiments is the weighted average of the gene expression of all the cell types in the tissue. In this framework, the expression of gene $j$ for the $i$-th patient can be modeled as a Gaussian distribution with mean parameter being the linear combination of the expression of gene $j$ in different cell types as follows:

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

with $K$ being the total number of cell types in the tissue, $\pi_{i,k}$ 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 $\sigma_j$ being the variance of the $j$-th gene. Reference-based models would consider $\mu_{k,j}$ as fixed with measurements derived from existing transcriptomic data from purified cells [4, 18, 34]; while reference-free models would estimate that from the data. A Bayesian model would specify prior distributions 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 $\{\mu_{k,j}\}$ [20].

# 3 Methods

## 3.1 Prior Specification

Let us assume that for each $k$-th cell type, there is a set $I_k$ of markers whose expression is upregulated in the $k$-th cell type compared to all others. We will use a flexible repulsive prior [20] in order to ensure that markers in set $I_k$ will have a "larger" mean in the $k$-th cell type compared to other cell types. Figure 1(A) provides a schematic of the proposed model. Let $\boldsymbol{\mu}_k$ be a $p$ dimensional vector containing the expression of $p$ markers in the $k$th cell type. Then, $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ is jointly modeled through the following multivariate prior distribution:

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.25.449763; this version posted August 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with $\xi_{k,j}$ and $\lambda_{k,j}$ being the mean and the variance parameter of the Gaussian prior distribution, $1(A)$ being an indicator function equal to 1 if A is satisfied and 0 otherwise, $\tau > 0$ is a scale parameter and $\eta > 0$ a positive integer controlling the rate at which the repulsive function approaches zero. This function is an extension of the repulsive function introduced by Petralia et al [20], and it approaches zero as the distance between mean parameters goes to zero and the upregulation of markers belonging to set $I_k$ in the $k$-th cell type is not satisfied. According to this function, markers 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 markers contained in set $I = \cup_{k=1}^{K} I_k$ will be assigned a repulsive prior; other markers will have a Gaussian prior distribution. 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 data. Depending on the cell types considered in the model, it might be problematic to find a set of markers upregulated in a particular cell type compared to all other cell types. For instance, the set of markers upregulated in Naive CD4 T Cells compared to Myeloid cells might be very different from the set of markers upregulated in Naive CD4 T Cells compared to Memory CD4 T Cells. In order to better leverage prior information, our flexible framework allows the user to specify cell type specific makers for each pair of cell types. For instance, assume that $I_{s>k}$ is the set of markers 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:

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

Standard choice for the parameters of the Gaussian prior is $\xi_{k,j} = 0$ and $\lambda_{k,j} = 1$; alternatively, those parameters might be chosen based on prior knowledge from existing databases and single-cell datasets. For ease of computation, we will not require $\{\pi_{i,k}\}_{k=1}^{K}$ to sum to 1. However, we will require these parameters to be defined on the unit interval $[0, 1]$. As prior specification, we will use a spike-and-slab prior [7] 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$(a_\gamma, b_\gamma)$. The spike component concentrates its mass at values close to zero, shrinking small effects to zero and inducing sparsity in the estimates $\{\pi_{i,k}\}$. Since some cell types will be more abundant (i.e., different from zeros) than others, the percentage of zero values, i.e., $w_k$, will vary across different cell types. 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.

## 3.2 Multi-omic Framework

BayesDeBulk can be utilized to perform the deconvolution by integrating gene expression and protein expression data as illustrated by Figure 1(B). In this case, each data type can be modeled via a BayesDeBulk model, with different data-specific models sharing the same set of cell type fraction parameters. Let $y_{i,j}$ and $z_{i,j}$ be the RNA and protein expression of marker $j$ for sample $i$. The proposed multi-omic framework models $y_{i,j}$ and $z_{i,j}$ as follows:

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

$$z_{i,j} \sim N(\theta^Z_{i,j}, \iota_j) \ \ , \ \ \theta^Z_{i,j} = \sum_{k=1}^{K} \pi_{i,k} \mu^Z_{k,j}$$

with $K$ being the total number of cell types, $\pi_{i,k}$ being the fraction of the $k$-th cell type for sample $i$, $\mu^Y_{k,j}$ being the expression of gene $j$ for the $k$-th cell type, $\mu^Z_{k,j}$ 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 type fractions, i.e. $\{\pi_{i,k}\}$. In this case, a repulsive prior will be placed on the mean parameters of both models, i.e., $\{\mu^Y_{k,j}\}$ and $\{\mu^Z_{k,j}\}$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.25.449763; this version posted August 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

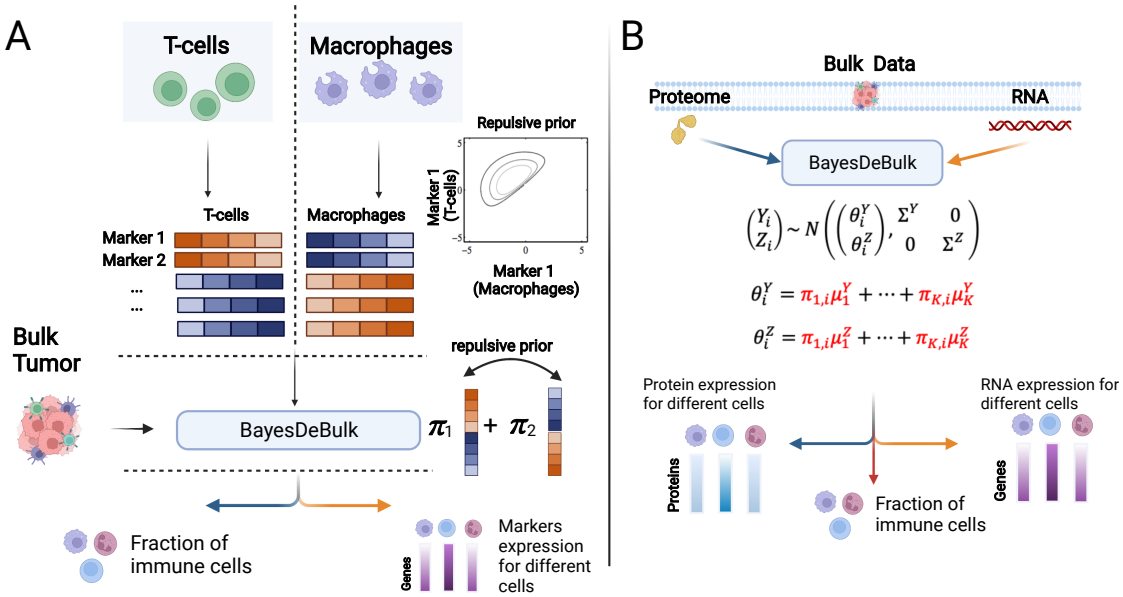


Figure 1: **Algorithm Schematic** (A) Bulk data is modeled as a linear combination of marker expression in different cell types. Given a list of markers expressed in each cell type, a repulsive prior is placed on the mean of marker expression in different cell types to ensure that cell type specific markers are upregulated in a particular component. (B) Multi-omic framework to estimate cell type fractions integrating global proteomic and RNAseq data. Given a list of cell-type specific markers, the algorithm returns the estimated protein/RNA expression for different cell types and cell-type fractions for different samples.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.25.449763; this version posted August 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

### 3.3 Full conditionals and posterior computation

Posterior computation of model parameters will be performed via Gibbs sampling [30]. Following Petralia et al [20], a latent variable $\rho$ 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:

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

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., $\pi_{i,k} \sim N_{[0,1]}(0, \gamma_k)$. The Gibbs sampler steps are summarized in Section 1 and 2 of the Supplementary Material.

## 4 Validation based on synthetic data

### 4.1 Data generated from a Gaussian model

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

BayesDeBulk was compared with Cibersort [18], Plier [15], xCell [1] and EPIC [25] 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 $\nu$ and $\sigma$. 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. 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 $1,000$. Different methods were implemented (i) assuming that all cell-type specific markers are known a priori and (ii) only 50% of cell-type specific markers are known a priori. This second scenario is more representative of real world applications, where only a proportion of cell-type specific markers is usually known. Cibersort and EPIC require as input a signature matrix containing the expression 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 (further information can be found in Section 3 of the Supplementary Material and Supplementary Figure 2).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.25.449763; this version posted August 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2 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 only 50% 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-type fractions for different synthetic data scenarios. A median correlation above 0.90 was observed for BayesDeBulk for all simulation scenarios involving $K = 10$ components. Among other deconvolution algorithms, Plier resulted in the best performance in terms of Pearson's correlation with a median correlation greater than 0.85 for all synthetic scenarios involving $K = 10$ cell types, although leading to the worst MSE. In addition, given the identifiability issue affecting factor models, for many synthetic scenarios, Plier was not able to map factors to a particular cell type (see Section 3 of Supplementary Material for further information). This identifiability issue is expected to be more pronounced in real-world applications, posing considerable challenges to Plier's implementation. As expected, the performance of all models decreased as more cell types were considered (K=20). Given their ability to estimate the expression of different markers for multiple cell types, BayesDeBulk and Plier were also compared in this regard. Figure 2 shows the performance of BayesDeBulk and Plier in estimating the mean of marker expression for different components. As expected, higher noise levels resulted in lower performance in terms of both correlation and MSE. The median Pearson's correlation between estimated and true values across replicates was above 0.80 for the simulations involving 10 cell types. Although the median correlation decreased substantially when the number of components increased to $K = 20$, it remained above 0.50 for different simulation scenarios. Additional synthetic data scenarios based on lower sample size $n = 50$ and different degree of prior knowledge can be found in Supplementary Figures 3, 4 and 5. Overall, we observe that the performance of BayesDeBulk is not affected by different degree of prior knowledge contrary to Cibersort, Epic and xCell.

# 5 Validation based on flow cytometry

In this section, the performance of BayesDeBulk is compared with Plier [15], Cibersort [18], xCell [1], CibersortX [19], EPIC [25] and MCP-counter [2] based on transcriptomic data from peripheral blood mononuclear cells. For this comparison, we used two public gene expression data based on influenza vaccination cohorts referred to as influenza cohort 1 [33] and influenza cohort 2 [3, 16, 17]. As additional cohort, we used the gene expression of a peripheral blood data cohort involving 20 patients [18]. For BayesDeBulk, Epic, Plier and Cibersort inference the LM22 signature matrix from Cibesort was considered. For MCP-counter and xCell estimation, their default signature was utilized. Further details on how the signature matrix was leveraged for different algorithms can be found in Section 4 of the Supplementary Material.

Figure 3 shows the Spearman's correlation and MSE between flow-cytometry data and estimates derived using different algorithms. BayesDeBulk resulted in an overall Spearman's correlation higher than all other algorithms for the influenza cohort 1. Cibersort and CibersortX models performed poorly in estimating the fraction of Monocytes. All algorithms performed well in estimating the fraction of B cells and CD8 T cells. For influenza cohort 2, BayesDeBulk was outperformed only by MCP-counter in terms of both Spearman's correlation and mean squared error. Again, Cibersort and CibersortX performed poorly in estimating Monocytes fractions. For both data sets, MCP-counter and BayesDeBulk resulted in the lowest MSE considering all cell types combined. Finally, for the peripheral blood data involving 20 samples, BayesDeBulk was outperformed only by Cibersort and CibersortX in terms of Spearman's correlation. Plier was not able to map any estimated factors to CD8 T cells; and therefore summary statistics for this cell type were not reported. The overall Spearman's correlation considering the

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.25.449763; this version posted August 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

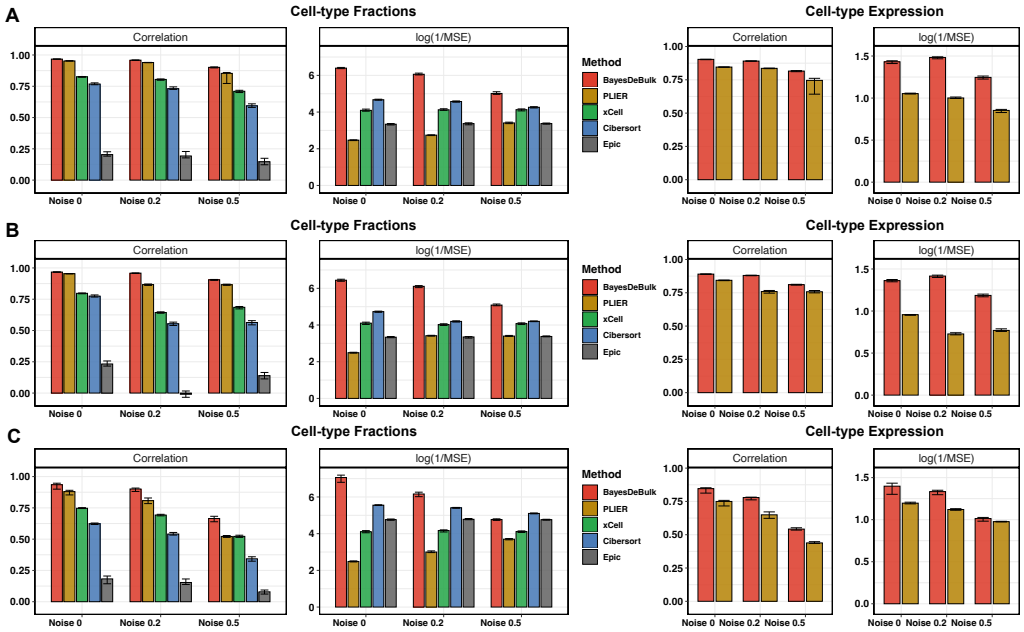


Figure 2: Pearson's correlation and mean squared error (MSE) between estimated values and true values over 30 replicates for BayesDeBulk (red), Cibersort (blue), xCell (green), Epic (gray) and Plier (gold). Barplots correspond to the median across different replicates while error bars to the interquartile range. For each simulation scenario, we report the correlation and MSE between the estimated cell-type fractions and the true values (left-hand panel) for all five algorithms, and between the estimated cell-type expression and true values (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 $\nu$ (noise).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.25.449763; this version posted August 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

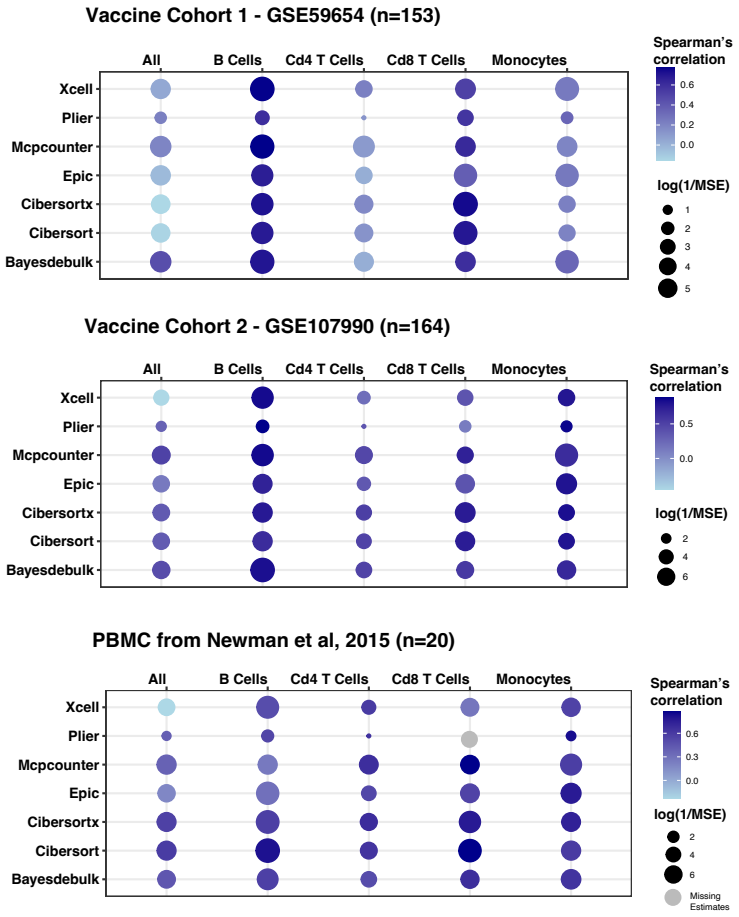


Figure 3: Spearman's correlation and MSE between estimated cell-type fractions and estimates based on flow-cytometry for BayesDeBulk, Cibersort, CibersortX, Plier, EPIC and MCP-counter for different datasets.

remaining cell types was lower than that of BayesDeBulk (spearman's correlation of 0.40 versus 0.45). For all three datasets, Plier was the algorithm resulting in the worst mean squared error. Since BayesDeBulk perform the estimation of both cell-type fractions and cell-type gene expression, we expect BayesDeBulk to perform less accurately for small number of samples compared to algorithms using a fixed signature matrix (e.g., Cibersort). However, as shown by this data example, BayesDeBulk is still among the top performers in estimating cell-type fractions for datasets involving a small number of samples (n=20).

# 6 Validation based on mixture of protein and gene expression from purified cells

In this section, we compare different algorithms based on mixture of gene and protein expression datasets from purified cells. For this experiment, we considered data from [13] 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. Considering the transcriptomic profiles of these immune cells, we obtained pseudo-bulk RNA data as explained in Section 5 of the Supplementary Material.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.25.449763; this version posted August 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Then, we considered data from Rieckmann et al [27] including proteomic profiles for the same set of immune cells. Mixed proteomic data was generated in a similar fashion as the transcriptomic profile, considering the same set of mixture proportions. BayesDeBulk was compared with Cibersort [18], Epic [25], Plier [15], xCell [1] and MCP-counter [2] in estimating immune cell-type fractions. Considering a sample size of $n = 50$, 30 replicate datasets were generated and the performance of the models were evaluated based on the 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 $1,000$. All algorithms were implemented using proteomic and transcriptomic data. Given the ability of BayesDeBulk to integrate both proteomic and RNAseq data, a multi-omic based learning for BayesDeBulk was also implemented. For BayesDeBulk, Epic, Plier and Cibersort inference the LM22 signature matrix from Cibersort was considered. For MCP-counter and xCell estimation, their default signature was utilized. Further details on how the signature matrix was leveraged for different algorithms can be found in Section 5 of the Supplementary Material.

Figure 4 shows the Spearman's correlation and mean squared error (MSE) between true and estimated cell-type fractions for different algorithms. Overall, BayesDeBulk resulted in the highest correlation and lowest MSE. For some cell types such as Neutrophils, the multi-omic based deconvolution was able to slightly outperform single-omic based deconvolutions revealing the advantage of a multi-omic based learning. As shown by Figure 4, Epic was the worst algorithm when performing the deconvolution based on proteomic data; resulting in a negative correlation for CD4 T cells and CD8 T cells. Although Plier was the algorithm with the highest correlation among competitors, it resulted in the highest MSE. Similarly to other data examples, for a large number of replicates ($\sim 50\%$), Plier could not map estimated factors to each one of the 6 immune cell types. Only replicates for which this mapping was possible were considered to produce summary statistics in Figure 4. As previously highlighted, this identifiability problem is an issue when applying Plier to real-world data.

# 7 Conclusion

We introduce BayesDeBulk, a new Bayesian method for the deconvolution of bulk data which can be applied to both gene-expression and protein-expression data or their integration. BayesDeBulk allows the simultaneous estimation of both cell-type fractions and marker expression for different cell types. As prior information, our proposed framework requires a set of cell-type specific markers which can be obtained from existing gene expression and protein expression from purified cells or single-cell experiments. BayesDeBulk models the bulk proteomic and transcriptomic data via a Gaussian distribution with mean being the linear combination of marker expression in different cell types. By leveraging existing prior knowledge on cell-type specific markers, a repulsive prior is placed on the mean of marker expression in different cell types to ensure that cell-type specific markers are upregulated in a particular component. This prior specification facilitates the identification and the labeling of the components corresponding to different cell types.

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

In addition, the estimation of markers' abundance of different cell types is very

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.25.449763; this version posted August 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

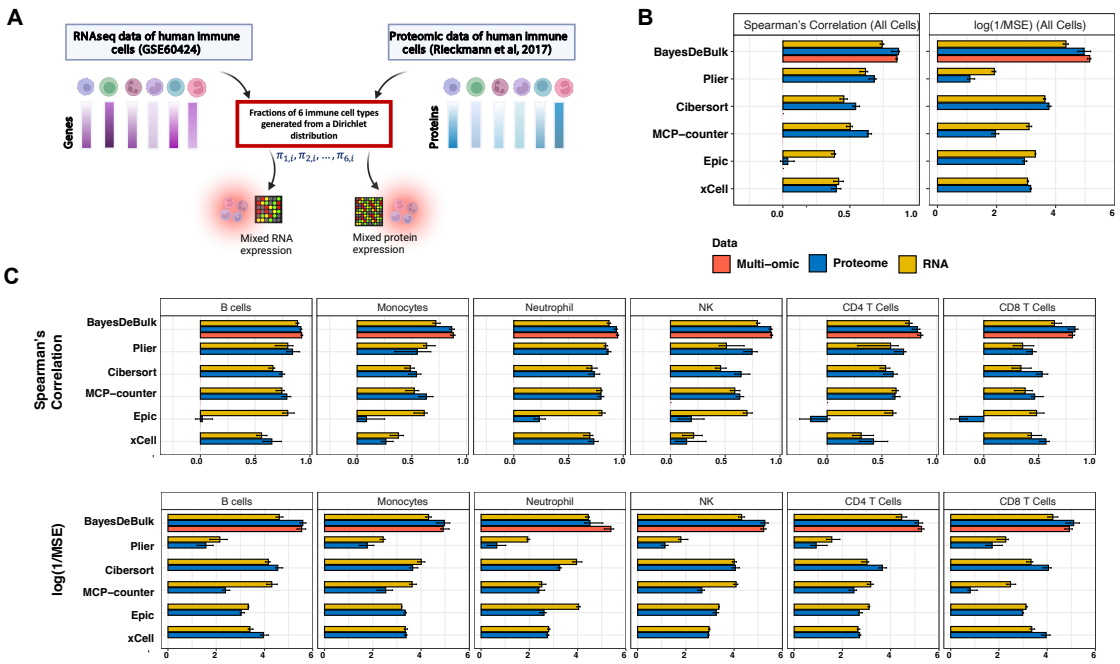


Figure 4: (A) Pseudo-bulk gene- and protein- expression were generated based on immune profiles from two publicly available data sets. The same set of mixture weights drawn from a Dirichlet distribution were considered in order to obtain pseudo-bulk data. (B-C) Spearman's correlation and mean squared error (log(1/x) scale) between true and estimated cell-type fractions for BayesDeBulk, Cibersort, EPIC, MCP-counter, Plier and xCell. Barplots correspond to median values, while error bars to interquartile range across 30 replicates. All algorithms were implemented based on proteomic (Pro) and transcriptomic (RNA) data. For BayesDeBulk, a multi-omic based deconvolution was also performed (Multi-omic).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.25.449763; this version posted August 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

important in order to perform differential expression analyses between adjacent normal and tumor tissues while accounting for tumor purity. 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 composition in tumor tissue. BayesDeBulk can be used to estimate the marker abundance in tumor cells by adding an extra component. This estimated tumor profile can be used to identify markers differentially expressed between tumor-cells and the adjacent normal tissue.

The performance of BayesDeBulk was compared with Cibersort, Epic, xCell, MCP-counter and Plier for different synthetic data examples. As shown, BayesDeBulk resulted in superior performance compared to other algorithms based on synthetic data generated from a Gaussian model or from a mixture of protein/gene expression of purified cells. In particular, we demonstrated that the multi-omic deconvolution resulted in superior performance compared to single-omic deconvolution, confirming the importance of multi-omic data integration. We then evaluated the performance of BayesDeBulk and other methods based on data with flow-cytometry measurements. We demonstrated that BayesDeBulk was among the top-performers in estimating cell-type fractions for different data sets.

Given that BayesDeBulk performs a simultaneous estimation of cell-type fractions and markers' expression in different cell types, the number of parameters to be estimated can quickly increase with the number of cell types. For this reason, we encourage the user to estimate a moderate number of cells ($K < 20$), especially when working with a moderate sample size ($n < 100$).

Besides proteomic and RNAseq data, BayesDeBulk could be easily extended to integrate other data types such as methylation profiles and single cell omics data measured for the same set of tumors. With the advancement of sequencing technology, it is more and more common to obtain multi-omic data for the same set of samples. In this framework, there is an urgent need of flexible algorithms able to integrate disparate omics data in order to better estimate the composition of the tumor microenvironment.

# 8 Data Availability

Gene expression data for the two influenza vaccination cohorts can be found in GEO GSE107990 and GSE59654. Flow cytometry data for both cohorts can be found in supplementary data of Monaco et al [16].

# 9 Acknowledgement and Funding

This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

# 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] 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.

[3] Christophe Carre, Glenn Wong, Vipin Narang, Crystal Tan, Joni Chong, Hui Xian Chin, Weili Xu, Yanxia Lu, Michelle Chua, Michael Poidinger, et al. Endoplasmic reticulum stress response and bile acid signatures associate with multi-strain seroresponsiveness during elderly influenza vaccination. *Iscience*, 24(9):102970, 2021.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.25.449763; this version posted August 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[4] 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*, pp. 243–259. Springer, 2018.

[5] David J Clark, Saravana M Dhanasekaran, Francesca Petralia, Jianbo Pan, Xiaoyu Song, Yingwei Hu, Felipe da Veiga Leprevost, Boris Reva, Tung-Shing M Lih, Hui-Yin Chang, et al. Integrated proteogenomic characterization of clear cell renal cell carcinoma. *Cell*, 179(4):964–983, 2019.

[6] 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.

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

[8] Michael A Gillette, Shankha Satpathy, Song Cao, Saravana M Dhanasekaran, Suhas V Vasaikar, Karsten Krug, Francesca Petralia, Yize Li, Wen-Wei Liang, Boris Reva, et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell*, 182(1):200–225, 2020.

[9] Richard W Griffiths, Eyad Elkord, David E Gilham, Vijay Ramani, Noel Clarke, Peter L Stern, and Robert E Hawkins. Frequency of regulatory t cells in renal cell carcinoma patients and investigation of correlation with survival. *Cancer Immunology, Immunotherapy*, 56(11):1743–1753, 2007.

[10] 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.

[11] 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.

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

[13] 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.

[14] Francesco Liotta, Mauro Gacci, Francesca Frosali, Valentina Querci, Gianni Vittori, Alberto Lapini, Veronica Santarlasci, Sergio Serni, Lorenzo Cosmi, Laura Maggi, et al. Frequency of regulatory t cells in peripheral blood and in tumour-infiltrating lymphocytes correlates with poor prognosis in renal cell carcinoma. *BJU International-British Journal of Urology*, 107(9):1500, 2011.

[15] 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.

[16] Gianni Monaco, Bernett Lee, Weili Xu, Seri Mustafah, You Yi Hwang, Christophe Carré, Nicolas Burdin, Lucian Visan, Michele Ceccarelli, Michael Poidinger, et al. Rna-seq signatures normalized by mrna abundance allow absolute deconvolution of human immune cell types. *Cell reports*, 26(6):1627–1640, 2019.

[17] Vipin Narang, Yanxia Lu, Crystal Tan, Xavier FN Camous, Shwe Zin Nyunt, Christophe Carre, Esther Wing Hei Mok, Glenn Wong, Sebastian Maurer-Stroh, Brian Abel, et al. Influenza vaccine-induced antibody responses are not impaired by frailty in the community-dwelling elderly with natural influenza exposure. *Frontiers in immunology*, p. 2465, 2018.

[18] 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.

[19] Aaron M Newman, Chloé B Steen, Chih Long Liu, Andrew J Gentles, Aadel A Chaudhuri, Florian Scherer, Michael S Khodadoust, Mohammad S Esfahani, Bogdan A Luca, David Steiner, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology*, 37(7):773–782, 2019.

[20] 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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.25.449763; this version posted August 3, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[21] Francesca Petralia, Nicole Tignor, Boris Reva, Mateusz Koptyra, Shrabanti Chowdhury, Dmitry Rykunov, Azra Krek, Weiping Ma, Yuankun Zhu, Jiayi Ji, et al. Integrated proteogenomic characterization across major histological types of pediatric brain cancer. *Cell*, 183(7):1962–1985, 2020.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] Reem Saleh and Eyad Elkord. Foxp3+ t regulatory cells in cancer: Prognostic biomarkers and therapeutic targets. *Cancer Letters*, 490:174–185, 2020.

[29] Shankha Satpathy, Karsten Krug, Pierre M Jean Beltran, Sara R Savage, Francesca Petralia, Chandan Kumar-Sinha, Yongchao Dou, Boris Reva, M Harry Kane, Shayan C Avanessian, et al. A proteogenomic portrait of lung squamous cell carcinoma. *Cell*, 184(16):4348–4371, 2021.

[30] Adrian FM Smith and Gareth O Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(1):3–23, 1993.

[31] 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.

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

[33] Juilee Thakar, Subhasis Mohanty, A Phillip West, Samit R Joshi, Ikuyo Ueda, Jean Wilson, Hailong Meng, Tamara P Blevins, Sui Tsang, Mark Trentalange, et al. Aging-dependent alterations in gene expression and a mitochondrial signature of responsiveness to human influenza vaccination. *Aging (Albany NY)*, 7(1):38, 2015.

[34] 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.

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

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