The Annals of Applied Statistics
2021, Vol. 15, No. 1, 391–411
https://doi.org/10.1214/20-AOAS1376

BAYICE: A BAYESIAN HIERARCHICAL MODEL FOR SEMIREFERENCE-BASED DECONVOLUTION OF BULK TRANSCRIPTOMIC DATA

BY AN-SHUN TAI[1,*], GEORGE C. TSENG[2] AND WEN-PING HSIEH[1,†]

[1]*Institute of Statistics, National Tsing Hua University,* [*]*anshuntai@nctu.edu.tw;* [†]*wphsieh@stat.nthu.edu.tw*
[2]*Department of Biostatistics, University of Pittsburgh, ctseng@pitt.edu*

Gene expression deconvolution is a powerful tool for exploring the microenvironment of complex tissues comprised of multiple cell groups using transcriptomic data. Characterizing cell activities for a particular condition has been regarded as a primary mission against diseases. For example, cancer immunology aims to clarify the role of the immune system in the progression and development of cancer through analyzing the immune cell components of tumors. To that end, many deconvolution methods have been proposed for inferring cell subpopulations within tissues. Nevertheless, two problems limit the practicality of current approaches. First, most approaches use external purified data to preselect cell type-specific genes that contribute to deconvolution. However, some types of cells cannot be found in purified profiles, and the genes specifically over- or under-expressed in them cannot be identified. This is particularly a problem in cancer studies. Hence, a preselection strategy that is independent from deconvolution is inappropriate. The second problem is that existing approaches do not recover the expression profiles of unknown cells present in bulk tissues when the reference set of purified cell-specific profiles is incomplete which results in biased estimation of unknown cell proportions. Furthermore, it causes the shift-invariant property of deconvolution to fail which then affects the estimation performance. To address these two problems, we propose a novel semireference-based deconvolution approach, BayICE which employs hierarchical Bayesian modeling with stochastic search variable selection. We develop a comprehensive Markov chain Monte Carlo procedure through Gibbs sampling to estimate proportions, expression profiles and signature genes for a set of known reference cell types as well as an unknown cell type. Simulation and validation studies illustrate that BayICE outperforms existing semireference-based deconvolution approaches in estimating cell proportions. We further show that BayICE is applicable to single-cell RNA-seq data. Subsequently, we demonstrate an application of BayICE in the RNA sequencing of patients with nonsmall cell lung cancer. The model is implemented in the R package "BayICE," and the algorithm is available for download.

1. Introduction.

Exploring the cellular components of heterogeneous tissues from their gene expression profiles is an essential work for revealing molecular mechanisms across different cell types. For instance, increasing evidence suggests that levels of tumor-infiltrating immune cells are associated with tumor progression, response to therapy and patient survival (Dieu-Nosjean et al. (2014); Fridman et al. (2012, 2017)). Thus, powerful technologies for single-cell isolation, such as laser microdissection and flow cytometry, have been employed to quantify the numbers of malignant and normal cells in tissue (Hu et al. (2016)). However, these physical approaches to isolating cells of interest at the gene expression level are costly and time consuming, resulting in drastically reduced biological-content yields. In contrast to

Received August 2019; revised April 2020.

Key words and phrases. Semi-reference-based deconvolution, hierarchical Bayesian, stochastic search variable selection, cell proportion, gene expression.

single-cell technologies, RNA-seq and microarrays yield bulk gene expression from hundreds of thousands of cells. In heterogeneous tissues, where more than one cell type is present, the expression profile from bulk RNA-seq or microarrays is from cell mixtures; thus, to correctly interpret these data, gene expression deconvolution approaches are required to recover cell type-specific expression and the distinct cellular proportions within complex tissues.

In the study of gene expression deconvolution, numerous computational and statistical approaches have been proposed to characterize cell subpopulations within tissues (Anghel et al. (2015); Becht et al. (2016); Gong et al. (2011); Li et al. (2016); Newman et al. (2015, 2019); Ogundijo and Wang (2017); Racle et al. (2017); Wang et al. (2019); Xie, Zhou and Xu (2018); Zhong et al. (2013)). Expression data from a heterogeneous tissue can be modeled as a linear combination of the distinct expression profiles of the cells present in that tissue, weighted by the corresponding cell fractions. These approaches can be grouped into one of three categories depending on whether they use a prior database of cell type-specific expression profiles in the deconvolution procedure: reference-free, reference-based and semireference-based methods. Reference-free approaches aim to directly perform expression deconvolution without cell type-specific references, and their most significant feature is that they estimate the relative cellular proportions and simultaneously disentangle their expression profiles. For instance, many studies have been leveraged on nonnegative matrix factorization to decompose mixed gene expression matrices into cell fractions and their corresponding expression profiles (Gaujoux and Seoighe (2012); Prassas et al. (2012)). Although reference-free models are valuable in the exploration of an uncharacterized cell population, such as tumor subclones (Xie, Zhou and Xu (2018)), relating the cellular components they identify to specific cell types of interest is difficult. Hence, the results of reference-free approaches are unable to clarify the association between a particular cell type and disease progression.

By contrast, reference-based methods incorporate external expression profiles of pure cell samples for deconvolution. For example, the analytical tool CIBERSORT successfully borrows cell type-specific information to predict the immune cell components in blood tissues and tumors through ν-support vector regression (ν-SVR) (Charoentong et al. (2017); Newman et al. (2015)). Due to increasing availability of single-cell RNA-seq data, MuSiC and CIBERSORTx are two of the most recent reference-based methods that use a single cell-derived signature matrix (Newman et al. (2019); Wang et al. (2019)). A fundamental assumption about reference-based models is that all types of cells present in the target tissues are included in the reference set, and the cellular proportions should sum up to one. Unfortunately, the pure expression profile of malignant cells, a key component in tumors, is a great challenge even for single-cell sequencing because of the high genetic heterogeneity of tumors. Hence, reference-based models can only derive the relative cell proportions concerning the reference set rather than the exact proportions concerning the microenvironment. Therefore, the relative cell proportions are not comparable across samples. To overcome this problem, TIMER adopts a series of deconvolution procedures to adjust the relative cell proportions with tumor purity which is the content of malignant cells in a tumor (Li et al. (2016)).

The abovementioned limitation forms the main incentive for developing semireference-based deconvolution approaches. In 2017, Racle et al. proposed a framework for estimating the proportion of immune and cancer cells (EPIC) for RNA-seq data (Racle et al. (2017)). EPIC applies least-squares regression with a nonnegativity constraint to the deconvolution problem and requires that the sum of all cell proportions in each tissue must be less than or equal to one. When the sum is not equal to one, one minus the sum of the estimated cell proportions represents the fraction of uncharacterized cells in a tissue that is not accounted for by the reference set; this number is interpreted as the malignant cell proportion in a bulk tumor.

Although semireference-based models demonstrate the advantages of incorporating cell-specific information and, simultaneously, extracting the uncharacterized cell types present in tissues, two problems behind these models should be addressed to complete the framework of gene expression deconvolution. First, signature gene selection is critical to the performance of gene expression deconvolution. In some studies the incorporation of a preselected signature gene set has successfully improved the accuracy of immune cell deconvolution (Chen et al. (2017, 2018); Newman et al. (2015); Racle et al. (2017)). However, the gene activities of a particular cell type usually vary across different tissue microenvironments. Hence, the use of a preselected gene set might cause the loss of data-dependent information and lead to less deconvolution power. The second problem concerns the natural characteristics of deconvolution. In a reasonable strategy for deconvolution, shifts in the mean level of reference samples and tumor samples should not change the estimation of cell proportions. We refer to this as the shift-invariant property for deconvolution. However, the constrained model implemented for EPIC does not maintain the shift-invariant property in deconvolution, and the estimation of cellular components is unstable because the sequencing depth changes across experiments. More specifically, the uncharacterized cell fractions estimated using the constrained least-squares approach are biased toward zero which will be demonstrated in our results.

Therefore, to address the aforementioned problems, we propose a new model based on a hierarchical Bayesian framework for intracellular component exploration. It is called BayICE and is a semireference-based approach. Under the Gaussian assumption, we first considered stochastic search variable selection (SSVS) for novel signature gene selection (George and McCulloch (1993)). The SSVS approach has been widely used in transcriptome analyses to select significant genes. For example, Ishwaran and Rao introduced a rescaled Bayesian model for selecting differentially expressed genes through multigroup microarray data (Ishwaran and Rao (2005a)). To the best of our knowledge, BayICE is the first attempt to incorporate the mechanism of feature selection for inferring the cellular components of bulk tissues when the reference set is incomplete. Although some reference-based methods can directly apply all genes to deconvolution by a weighting scheme (Wang et al. (2019)) or a matrix factorization technique (Sun, Sun and Yang (2019)) without signature gene preselection, the existing semireference-based methods still require preselecting a signature gene set. Moreover, we claim that the BayICE model possesses the shift-invariant property of deconvolution which yields unbiased estimates of cellular proportions. The model with the shift-invariant property further guarantees that it can recover the expression profiles of uncharacterized cells using posterior mean inference. For the purpose of inference, we applied Gibbs sampling and the Metropolis–Hastings as the sampling procedure in the estimation. In brief, BayICE performs cellular component estimation, uncharacterized cell expression profile estimation and a novel strategy for signature gene selection.

The remainder of this paper is organized as follows. Section 2 introduces the deconvolution in gene expression and states the shift-invariant property. Section 3 introduces the statistical modeling of BayICE for gene expression deconvolution and proposes a Markov chain Monte Carlo (MCMC) algorithm for simulating the posterior distributions of parameters. To assess the model's performance, Section 4 presents simulation studies that investigate gene expression deconvolution, gene selection and model robustness compared to two existing methods. Section 5 presents applications to three real datasets, including one single-cell RNA sequencing dataset, where underlying true cell proportions are known and performance can be benchmarked. Section 6 describes application of BayICE to 199 nonsmall cell lung cancer RNA-seq samples and exploration of the cell components present in the microenvironments of lung tumors. Finally, Section 7 provides final discussion and conclusions.

2. Deconvolution.

The deconvolution in gene expression can be formalized as an optimization problem in which the parameter of interest is the cellular proportion $W =$

$(w_1, \dots, w_K)'$, and the estimates of W are obtained by

$$\hat{W} = \underset{W}{\operatorname{argmin}}\, D\left(\begin{pmatrix} y_1 \\ \vdots \\ y_G \end{pmatrix}, \begin{pmatrix} B_{11} & \cdots & B_{1K} \\ \vdots & \ddots & \vdots \\ B_{G1} & \cdots & B_{1K} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_K \end{pmatrix}\right), \tag{2.1}$$

where y_g is the gene expression of gene g, B_{gk} is the expression of the kth cell type in gene g and D is the distance metric. In general, D is the Euclidean distance. As mentioned above, the reference-free deconvolution approaches assume the cell type-specific expression (B_{gk}) is unobserved, and the reference-based methods, by contrast, require B_{gk} as input for the optimization problem. Additionally, the semireference-based approaches allow that one of the cell types is absent in the reference. We now define the shift-invariant property to characterize different deconvolution models.

DEFINITION 1 (Shift-invariant property). Let Y be the mixed expression from bulk tissue and B be the cell type-specific expression matrix. If the cellular proportion estimate of a deconvolution method M is invariant when the expression distribution shifts with a constant (location parameter), then M has the *shift-invariant property*. That is, M satisfies

$$\underset{W}{\operatorname{argmin}}\, D(Y, BW) = \underset{W}{\operatorname{argmin}}\, D\big(Y + c, (B + c)W\big)$$

for any constant c, where D is the distance metric used by M.

The shift-invariant property in Definition 1 is essential for evaluating the deconvolution methods, especially in the genomic study. Since the protocol of a gene expression experiment is typically designed for each study, the mean read depth varies across experiments. More specifically, the different experimental protocols cause the location parameters of expression distributions to change. If the location parameter affects the estimation of the same composition, then it is not reasonable to compare results across studies. However, the shift-invariant property guarantees that a deconvolution method with this property can estimate the proportions precisely when the location parameter is changed, and, thus, the between-study comparison is valid. In the Supplementary Material (Tai, Tseng and Hsieh (2021)), we have shown that the reference-free and reference-based deconvolution approaches follow the shift-invariant property. By contrast, EPIC adopted the inequality-constrained optimization method to derive cell proportions lacks the shift-invariant property, and, hence, the following simulation result reveals that the estimates of EPIC are biased. Moreover, the shift-invariant property reveals that the estimation bias increases as the baseline expression level increases. In the Supplementary Material (Tai, Tseng and Hsieh (2021)), we provided a toy example to demonstrate how the correlation between the real proportion and estimated proportion varies as the baseline expression level increases if the method does not have the shift-invariant property.

To address the issue of shift-invariant property for the semireference-based deconvolution approaches, we proposed a Bayesian deconvolution model which is more robust to the change in the location parameter. The Bayesian hierarchy architecture facilitates the construction of equality-constrained objective function for the semireference-based deconvolution problem via likelihood approach. The proof details in the Supplementary Material (Tai, Tseng and Hsieh (2021)), and the model construction will be detailed in the next section.

3. BayICE deconvolution model. In this section, we present the proposed hierarchical Bayesian deconvolution model for intracellular component exploration with novel signature gene selection. Figure 1 provides a graphical representation of the BayICE hierarchical

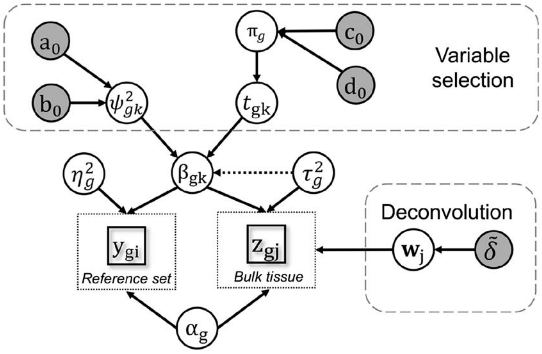

FIG. 1. *Bayesian hierarchical model of BayICE. The square symbols represent observed data from the reference set and bulk tissues. The white circles indicate priors, and the grey circles are hyperparameters.*

model. We first describe the input data and establish the statistical modeling for reference samples and tumor samples. Subsequently, a stochastic search method, the Bayesian false discovery rate and an inflation factor are introduced for signature gene selection. Finally, we adopt the Gibbs sampling approach and the Metropolis–Hastings approach to develop a comprehensive sampling procedure.

3.1. *Input data and normalization*. BayICE is a statistical framework designed for gene expression deconvolution. We first assume that the gene expression data are available from two sets of samples, namely, heterogeneous tissues with a given clinical condition and a reference set consisting of several groups of samples with pure cell types. Although BayICE is a quantitative model-based approach, microarray and RNA-seq expression data are both legal inputs for BayICE. For the microarray data we implement a locally weighted scatterplot smoothing algorithm (Yang et al. (2002)) for normalization. For the read count RNA-seq data, we recommend two different strategies for normalization. The first is to use new gene expression units, called transcripts per million (TPMs), which were proposed by Wagner, Kin and Lynch in 2012. The main feature of TPM normalization is to make the sum of all TPMs equal across samples which facilitates fair comparisons between samples. For public data, TPM normalization approach is not always available because the information of sequencing depth or gene length required for TPM could be missing. Thus, we adopt subsampling normalization as the second strategy. Subsampling normalization applies a binomial sampler to resample the read count, and it aims to maintain internal associations between genes and can simultaneously adjust external variance between samples. The details of how we implement the two normalization approaches are provided in the Supplementary Material (Tai, Tseng and Hsieh (2021)). After normalization by the first strategy or the second strategy, we consider a log transformation by $\log(\text{count} + 1)$. The log transformation of count data has been widely applied in RNA-seq studies. Furthermore, the batch-specific variation arises from aligning different datasets for deconvolution. In our analysis of the real application, the quantile normalization was implemented to remove platform-based bias from RNA-seq data. Additionally, one systematic review has tested 14 popular batch correction algorithms designed for single RNA-seq data, and they concluded that every batch-effect removal algorithm has its advantages and limitations with no superior method (Tran et al. (2020)). Those batch correction algorithms are available options for the preprocess of data before deconvolution.

3.2. *Statistical modeling*. The problem of gene expression deconvolution can be formulated as a system of linear equations that describes the expression of a given gene in a bulk tissue as the weighted sum of the expression values from multiple cell types present in the tissue. To maximize the deconvolution power, BayICE incorporates a reference set comprising cell-specific expression profiles into the inference of cellular components in bulk tissues. In this study the reference set contains various types of immune cells, such as T cells, B cells, natural killer cells, monocytes, neutrophils and normal tissue cells. In addition to these nonmalignant cells, malignant cells are a major cellular component of tissues in cancer deconvolution studies. Unfortunately, the high genetic heterogeneity of malignant cells hinders the possibility of constructing a predefined cancerous cell expression profile that can be applied to every bulk sample. To address this problem, BayICE takes the advantages of the flexibility of hierarchical Bayesian inference to extract malignant cell profiles directly from bulk tissues. Moreover, BayICE infers cellular proportions with respect to the reference cell types and unknown cell types of each tissue by integrating with a Bayesian signature gene selection approach.

We first introduce the statistical model for the reference set consisting of N purified samples with K cell types. We denote an observation in this set as y_{gi} which is the normalized value for gene-g ($g = 1, \dots, G$) in the ith purified sample ($i = 1, \dots, N$). The N purified samples belong to K nonmalignant cell types that exist in the bulk tissues. We introduce a binary vector variable $x_{i.} = (x_{i1}, \dots, x_{iK})$ to represent the cell type for the ith purified sample. More specifically, if the ith purified sample belongs to cell type k, then $x_{ik} = 1$ and $x_{it} = 0$ for all $t \neq k$. BayICE assumes that y_{gi} follows the Gaussian distribution with a mean level μ_{gi} and a gene-specific variance denoted by $\sigma^2_{y,g}$. For the purpose of signature gene identification, we consider the mean structure μ_{gi} to comprise a gene-specific baseline (α_g) and cell type-specific effects (β_{gk}). Thus, the modeling of reference data can be written as

$$(y_{gi} | \alpha_g, \{\beta_{gk}\}, \{x_{ik}\}, \sigma^2_{y,g}) \sim N(\mu_{gi}, \sigma^2_{y,g}) \tag{3.1}$$
$$\text{where } \mu_{gi} = \alpha_g + \sum_{k=1}^{K} x_{ik}\beta_{gk}.$$

To search for signature genes that exhibit a differential effect among cell types, we define a baseline cell type against which changes in expression levels are measured. We set β_{g1} to be 0 for all gene-g.

The cell type-specific effects ($\beta_{g1}, \dots, \beta_{gK}$) in (3.1) are shared in constructing the mean structure of gene expression in bulk tissues. It has been observed in the literature that gene expression of a certain cell type changed when it went through cell sorting (Richardson, Lannigan and Macara (2015); van den Brink et al. (2017)). To accommodate the effect of the changes induced by cell separation, we adopt the joint modeling of $\beta_{g1}, \dots, \beta_{gK}$ for pure cell samples and bulk samples.

For the model of the bulk tissue, it is a linear combination of K nonmalignant cell-specific expression profiles and malignant cell gene expression. The expression of gene-g from the jth bulk tissue ($j = 1, \dots, M$) is denoted by z_{gj}. The mean of z_{gj}, called υ_{gj}, can be modeled as the weighted sum of nonmalignant cell-specific effects $\{\beta_{gk}\}$ and one malignant cell-specific effect β_{g0} through a linear regression, given by

$$(z_{gj} | \alpha_g, \{\beta_{gk}\}, \{w_{jk}\}, \sigma^2_{z,g}) \sim N(\upsilon_{gj}, \sigma^2_{z,g}), \tag{3.2}$$
$$\upsilon_{gj} = \alpha_g + \sum_{k=1}^{K} w_{jk}\beta_{gk} + w_{j0}\beta_{g0}.$$

The weights, w_{jk}, are the proportions of expression attributable to normal cell type k in the jth tumor, and the weight, w_{j0}, represents tumor purity which is the percentage of malignant cells in a tumor tissue. Notably, BayICE can be used to explore not only tumors but also other noncancerous tissues. For noncancerous tissues, w_{j0} can refer to the proportion of one unknown cell type that is uncharacterized by the reference set. In our model a natural constraint for these cell proportions is that the sum of weights across cell types should be one (i.e., $\sum_{k=1}^{K} w_{jk} + w_{j0} = 1$ for each j). Furthermore, to characterize the gene expression pattern in malignant cells, we introduce a tumor-specific parameter β_{g0} to represent the effect size of gene-g in cancer. In a real application, multiple unknown cell types or cancerous cell types might be present in a bulk tissue. In this case, BayICE treats these uncharacterized cell types as a whole, and w_{j0} represents the proportion of unknown class in a bulk tissue.

3.3. *Novel gene selection using the SSVS approach.* Identifying signature genes that are expressed in a particular cell type is essential to the success of expression deconvolution. Although a preselected signature gene set could be easily applied to data analysis, the application of external signature genes could lose data-dependent information for deconvolution. Thus, we incorporate the stochastic search variable selection (SSVS) approach into our Bayesian deconvolution model for integrating expression deconvolution with novel signature gene selection. The SSVS approach, introduced by George and McCulloch (1993), specifies a spike-and-slab mixture prior which uses data to extract the potential features of the true model by inferring posterior probability. The spike component, which concentrates its mass at values close to zero, shrinks small effects to zero, whereas the slab component spreads its mass over a wide range of possible values for the effect size.

The proposed prior structure of BayICE on effect size exhibits a bimodal distribution on the variance of the coefficients that result in a spike-and-slab type prior on the effects themselves (Ishwaran, Kogalur and Rao (2010); Ishwaran and Rao (2005a, 2005b)). For each effect size β_{gk}, the prior structure is given by

$$\begin{aligned} (\beta_{gk}|\psi_{gk}^2, \pi_k) &\sim N(0, \psi_{gk}^2 t_{gk}), \\ (t_{gk}|\pi_k) &\sim \pi_k I_1(t_{gk}) + (1-\pi_k) I_c(t_{gk}), \\ (\psi_{gk}^2|a_0, b_0) &\sim \operatorname{Inv}\Gamma(a_0, b_0), \\ (\pi_k|c_0, d_0) &\sim \operatorname{Beta}(c_0, d_0), \end{aligned} \tag{3.3}$$

where $I_c(t_{gk})$ is a mass function that is 1 at $t_{gk} = c$ and 0 everywhere else. We set the value c as a small positive number in this study, such as $c = 10^{-5}$, and, thus, the random variable t_{gk} is 1 with probability π_k and close to zero with probability $1 - \pi_k$. When the gth gene is differentially expressed between the kth cell type and the other types, β_{gk} is more likely generated from the slab component and t_{gk} equals one. By contrast, $t_{gk} = c$ indicates that the gth gene is irrelevant to cell types, and its effect size is from the spike component. The hypervariance ψ_{gk}^2 is sampled from an inverse gamma with two given hyperparameters, a_0 and b_0. Following Ishwaran and Rao (2005a, 2005b), a_0 and b_0 are set as 5 and 50, respectively. The proportion of genes differentially expressed in cell type k is controlled by π_k, and we assume that π_k follows a beta distribution with $c_0 = 0.1$ and $d_0 = 0.1$ which is a noninformative prior for the binomial model.

To borrow variance information across samples, we modify the variance structure of effect size β_{gk} in (3.3) by considering the gene-specific variance $\sigma_{z,g}^2$ of mixture samples in (3.2), and the modified prior structure for β_{gk} is given by

$$\begin{aligned} (\beta_{gk}|\psi_{gk}^2, t_{gk}, \sigma_{z,g}^2) &\sim N(0, \psi_{gk}^2 t_{gk} \sigma_{z,g}^2), \\ (t_{gk}|\pi_k) &\sim \pi_k I_1(t_{gk}) + (1-\pi_k) I_c(t_{gk}), \\ (\psi_{gk}^2|a_0, b_0) &\sim \operatorname{Inv}\Gamma(a_0, b_0), \\ (\pi_k|c_0, d_0) &\sim \operatorname{Beta}(c_0, d_0). \end{aligned} \tag{3.4}$$

The role of the gene-specific variance $\sigma_{z,g}^2$ that appears in (3.4) can be intuitively interpreted as the baseline in the feature selection procedure. The modified prior structure considers the trade-off between the value of effect size and gene-specific variance to facilitate the establishment of feature selection.

3.4. *Bayesian false discovery rate.* In frequentist approaches to the test multiplicity problem, controlling the false discovery rate (FDR) has been widely applied to more adequately control genome-wide false positives. Whittemore in 2007 introduced a Bayesian FDR associated analogously with the frequentist FDR (Whittemore (2007)) as follows:

$$\varnothing_k(r) = \frac{\sum_{g=1}^{G} P_{gk}(H_0|Y, Z) D_{gk}(r)}{\sum_{g=1}^{G} D_{gk}(r)}, \tag{3.5}$$

where $P_{gk}(H_0|Y, Z)$ is the posterior probability that gene-g is not associated with cell type k (H_0) given observation (Y, Z) and $D_{gk}(r)$ is the rejection rule defined by $I(P_{gk}(H_0| Y, Z) < r)$. The tuning parameter r can be adjusted to control the Bayesian FDR at a certain α level. In the following simulations and applications the Bayesian FDR is used to address the multiplicity problem.

3.5. *Inflation factor.* In a Bayesian framework the influence of priors on posterior always vanishes as the sample size increases. This phenomenon limits Bayesian variable selection because the mechanism of such selection requires an effective prior setting. To overcome this disadvantage, Ishwaran and Rao (2005a, 2005b) proposed a rescaling approach to enable estimation invariant to the sample size by setting the prior as a function of the sample size. They applied data rescaling to the gene selection framework for multigroup microarray data. Furthermore, to achieve invariance to sample size, Ishwaran and Rao performed a sample size-related transformation of gene expression through multiplication by the global inflation factor

$$\sqrt{(\text{total sample size})/(\text{estimate of total variance})}.$$

Multiplication by this global inflation factor has been shown to ensure that the prior has a nonvanishing effect. Hence, following the concept of data transformation, we rescale the observations in our reference set and bulk tissue set as follows:

$$y_{gi}^* = \sqrt{\frac{N+M}{\hat{S}_g^2}} y_{gi} \quad \text{and} \quad z_{gj}^* = \sqrt{\frac{N+M}{\hat{S}_g^2}} z_{gj}, \tag{3.6}$$

where $\hat{S}_g^2 = \frac{1}{N-K} \sum_i \sum_k (y_{gi} x_{ik} - \overline{y}_{gk})^2$ is an unbiased estimator of σ_g^2 calculated from the reference set. Although we assume that the variance σ_g^2 of reference data $\{y_{gi}\}$ is shared with tissue data $\{z_{gj}\}$, the calculation of an unbiased estimator using both $\{y_{gi}\}$ and $\{z_{gj}\}$ data is a difficult task because of the convolution structure in $\{z_{gj}\}$. Note that the multiplier in (3.6) is a gene-specific inflation factor rather than the abovementioned global inflation factor since the inflation factor is composed of the total sample size and a gene-related variance. Therefore, the use of gene-specific factors can simultaneously achieve sample size invariance and gene-scale consistency.

After rescaling, the corresponding distributions for the transformed data $\{y_{gi}^*\}$ and $\{z_{gj}^*\}$ are modified as follows:

$$\begin{aligned} (y_{gi}^* | \alpha_g, \{\beta_{gk}\}, \{x_{ik}\}, \eta_g^2) &\sim N(\mu_{gi}, (N+M)\eta_g^2), \\ (z_{gj}^* | \alpha_g, \{\beta_{gk}\}, \{w_{jk}\}, \tau_g^2) &\sim N(\upsilon_{gj}, (N+M)\tau_g^2). \end{aligned} \tag{3.7}$$

The new variances of the transformed data are adjusted as sample size-related parameters, and this adjustment can be interpreted as a penalization shrinkage effect of the posterior mean. After the adjustment with inflation factors, the variances of y_{gi}^* and z_{gj}^* are asymptotically equal to $N+M$. For the purpose of variable selection, we introduce two further parameters η_g^2 and τ_g^2 for variances in (3.7) to keep the flexibility of modeling.

3.6. *MCMC sampling procedure.* Next, based on the transformed data $\{y_{gi}^*\}$ and $\{z_{gj}^*\}$, we complete the structure of the hierarchical Bayesian model in BayICE and then establish a sampling procedure to achieve signature gene selection and cell component inference. Following the abovementioned specification, the BayICE model is given by

$$\begin{aligned}
&(y_{gi}^*|\alpha_g,\{\beta_{gk}\},\{x_{ik}\},\eta_g^2)\sim N(\mu_{gi},(N+M)\eta_g^2), \quad i=1,\dots,N, g=1,\dots,G,\\
&\mu_{gi}=\alpha_g+\sum_{k=1}^{K}x_{ik}\beta_{gk},\\
&(z_{gj}^*|\alpha_g,\{\beta_{gk}\},\{w_{jk}\},\tau_g^2)\sim N(\upsilon_{gj},(N+M)\tau_g^2), \quad j=1,\dots,M, g=1,\dots,G,\\
&\upsilon_{gj}=\alpha_g+\sum_{k=1}^{K}w_{jk}\beta_{gk}+w_{j0}\beta_{g0},\\
&\boldsymbol{w}_j=(w_{j1},\dots,w_{jK},w_{j0})\sim \text{Dirichlet}(\delta,\dots,\delta), \quad j=1,\dots,M,\\
&(\alpha_g,\eta_g^2,\tau_g^2)\sim f(\eta_g^2,\tau_g^2)\propto\frac{1}{\eta_g^2}\times\frac{1}{\tau_g^2},\\
&(\beta_{gk}|\psi_{gk}^2,t_{gk},\tau_g^2)\sim N(0,\psi_{gk}^2t_{gk}\tau_g^2), \quad g=1,\dots,G,k=1,\dots,K,\\
&(t_{gk}|\pi_k)\sim\pi_kI_1(t_{gk})+(1-\pi_k)I_c(t_{gk}),\\
&(\psi_{gk}^2|a_0,b_0)\sim \text{Inv}\,\Gamma(a_0,b_0),\\
&(\pi_k|c_0,d_0)\sim\text{Beta}(c_0,d_0),
\end{aligned}\tag{3.8}$$

where hyperparameters are specified as $a_0=5$, $b_0=50$, $c_0=1$, $d_0=1$ and $\delta=1$ in this study.

Subsequently, based on the hierarchical prior setting, we apply Gibbs sampling and the Metropolis–Hastings approach to simulate the posterior value from

$$(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{T},\boldsymbol{\pi},\boldsymbol{W},\boldsymbol{\eta},\boldsymbol{\tau}^2,\boldsymbol{\psi}|\boldsymbol{Y}^*,\boldsymbol{Z}^*,\mathbf{X}),$$

where $\boldsymbol{\alpha}=\{\alpha_g\}$, $\boldsymbol{\beta}=\{\beta_{gk}\}$, $\boldsymbol{T}=\{t_{gk}\}$, $\boldsymbol{\pi}=\{\pi_k\}$, $\boldsymbol{W}=\{\boldsymbol{w}_j\}$, $\boldsymbol{\eta}=\{\eta_g^2\}$, $\boldsymbol{\tau}^2=\{\tau_g^2\}$, $\boldsymbol{\psi}=\{\psi_{gk}^2\}$, $\boldsymbol{Y}^*=\{y_{gi}^*\}$, $\boldsymbol{Z}^*=\{z_{gj}^*\}$ and $\mathbf{X}=\{x_{ik}\}$. The Gibbs sampler in BayICE works are shown in the Supplementary Material (Tai, Tseng and Hsieh (2021)). After a number of iterations, we specify the burn-in period and the thinning interval to obtain the posterior distribution of the parameters and then perform an analysis using the posterior mean or posterior median. In BayICE the default length of burn-in is 0.6 times the number of iterations, and the thinning interval has a length of three iterations. The convergence property is discussed in the Supplementary Material (Tai, Tseng and Hsieh (2021)).

4. Simulation study.

This section presents simulation results based on synthetic datasets to benchmark the performance of BayICE. We consider two types of data: array-based data, which is generated by a normal simulator, and sequencing-based data which can be produced by a multinomial simulator or negative binomial simulator. We mainly use the multinomial simulator to demonstrate the estimation of cellular proportions, detection of signature genes and the ability to recover the expression profile of unknown or malignant cells. Furthermore, we apply three different simulators to demonstrate the robustness of BayICE, and the results of the robustness study are shown in the Supplementary Material (Tai, Tseng and Hsieh (2021)).

4.1. *Multinomial simulator settings.* We first perform a sequencing-based simulation using a multinomial simulator for expression deconvolution. Thus, we consider a scenario in which five distinct cell subpopulations are present in a tissue, and one cell type is absent from our reference set. This simulation includes 5000 genes, with 300 genes designed as cell type-related genes. These 300 genes are divided randomly into five disjoint groups, and the genes assigned to a particular group are associated with one cell type. For each cell type in a reference set, we have 20 replicates; hence, the reference set includes 80 samples. In the reference set an 80 × 4 matrix of binary variables $X = \{x_{ik}\}$ is used to record the cell type of samples. Furthermore, in the bulk dataset we simulate the expression data of 90 mixed samples with different cell proportions. The simulation procedure is described as follows.

To account for the fact that the expression level varies across genes, we simulated data according to a set of real data from purified samples. We collected 19 RNA-seq samples of normal lung tissues from the Gene Expression Omnibus database with accession number GSE81089 (Mezheyeuski et al. (2018)) and then took the average gene expression across 19 samples to obtain the baseline a_g^{sim} of the gth gene. Among 17,775 genes we randomly picked 5000 genes for our simulation study. Based on each gene-specific expression level, we define the cell-specific effect as follows:

$$b_g^{\text{sim}} = (NES)^{c_g^{\text{sim}}}(-NES)^{1-c_g^{\text{sim}}} a_g^{\text{sim}},$$

where c_g^{sim} is a binary variable, which is 1 for upregulated status and 0 for downregulated status and *NES* is the normalized effect size set as one of the numbers $\{0.1, 0.2, \dots, 0.6\}$. The number $\{c_g^{\text{sim}}\}$ is randomly generated with a probability of 0.5 for 0 and 1. We further sample a series of values from Uniform(0.9, 1.1), called $\{e_i^{\text{sim}}\}$, as sample-specific effects because of the sample heterogeneity. Finally, we generate cellular proportions $\{w_{j1}^{\text{sim}}, \dots, w_{jK}^{\text{sim}}, w_{j0}^{\text{sim}}\}$. To explore the effect of unknown cell content on the model performance, we assign a fixed number to w_{j0}^{sim}, and the remaining components $\{w_{j1}^{\text{sim}}, \dots, w_{jK}^{\text{sim}}\}$ are generated from the Dirichlet distribution with parameters (1,…,1). Because of the sum-to-one constraint on cellular proportions, $\{w_{j1}^{\text{sim}}, \dots, w_{jK}^{\text{sim}}, w_{j0}^{\text{sim}}\}$ should be normalized by dividing their sum.

The mean expression structure is determined for both purified and mixed samples according to the abovementioned parameter settings. Notably, the use of a multinomial model in our simulation is to simulate the sequence alignment procedure that maps the reads against the reference sequence. The number of trials in the multinomial model is related to the total reads, and we set it as 5×10^6; in other words, the average read depth is designed as 1000. The probability of the multinomial model controls the expression levels across genes, and, therefore, we use the relative mean expression as the probability value. The complete sampling procedure details in the Supplementary Material (Tai, Tseng and Hsieh (2021)).

4.2. *Assessing the inference of deconvolution.* To assess the performance of BayICE deconvolution, we include two semireference-based approaches for comparison: EPIC and nonnegative least-squares (NNLS). NNLS is a general approach for solving the constrained least-squares problem where the coefficients are not allowed to become negative. We modify the NNLS approach by restricting the sum of coefficients to less than one for incomplete reference data deconvolution. We first examine the cellular proportion estimations obtained using

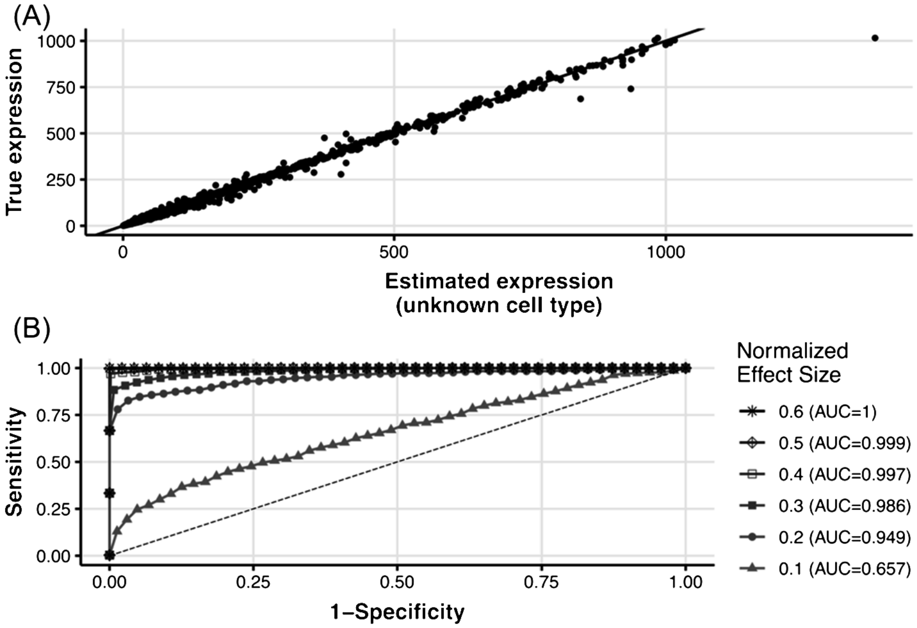

FIG. 2. *Results of unknown profile estimation and marker gene selection.* (*A*) *Scatter plot of gene expression of the unknown cell type between the truth and estimation.* (*B*) *ROC curves to evaluate the gene selection of BayICE under different normalized effect sizes. The diagonal dashed line represents the ROC curve for a random guess.*

EPIC, NNLS and BayICE. Notably, EPIC and NNLS both require an external step to identify signature genes before deconvolution. In this case we applied the marker genes identified by BayICE into EPIC and NNLS for a fair comparison.

According to the simulation setting, we generate 90 bulk samples per simulation in which the unknown cell proportions vary from 0.1 to 0.9. A large unknown cell proportion value indicates that the corresponding tissue is highly heterogeneous, such as in tumors; by contrast, a small proportion simulates the microenvironment of normal tissues in which most of the cell types can be purified. Additionally, we evaluate the performance of each model under different normalized effect sizes of marker genes. Moreover, BayICE can recover the underlying expression profile of the unknown cell type using the posterior mean. Because EPIC and NNLS cannot infer unknown expression profiles, we directly compare the estimation of uncharacterized cell profiles with the true mean expression. The results are shown in Figures 2 and 3.

We first evaluate the estimation of gene expression profiles for the unknown cell type and gene identification from BayICE in Figure 2. Figure 2(A) is a scatter plot between the true expression of the unknown cell type and the estimated expression. The results reveal the correlation between the real and estimated values to be greater than 0.98 which implies that BayICE can recover uncharacterized cell expression when one cell type is absent from the reference set. In Figure 2(B) we adopt the receiver operating characteristic (ROC) curve to quantify the results of gene identification under different normalized effect sizes which represent the strength of cell type-specific activity in marker genes. The performance in terms of area under the curve (AUC) is significantly improved with the increase of the effect size, and, more specifically, the AUC exceeds 0.94 when the effect is larger than 0.2. Subsequently, we apply root-mean-square error (RMSE) to quantify the accuracy of cell proportion estimation from BayICE, EPIC and NNLS. Figure 3(A) illustrates the changes in RMSE of all estimates with the increase of normalized effect size under the settings of unknown cell proportion = 0.1, 0.5 and 0.9. In Figure 3(B) we fix the normalized effect size at 0.1, 0.3 and 0.6 and

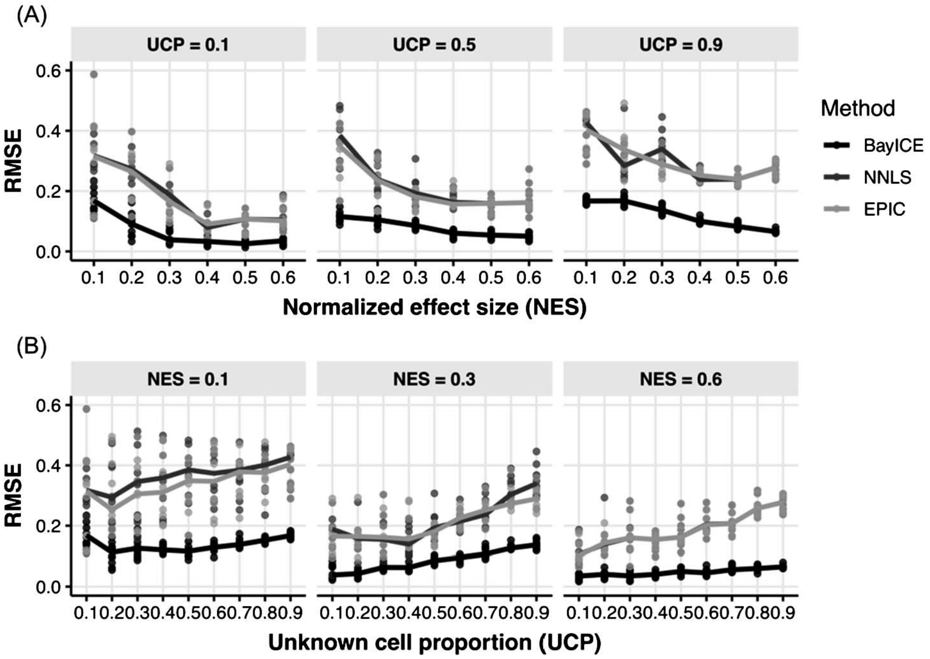

FIG. 3. *Deconvolution results of cell proportion estimation. Root-mean-square error* (*RMSE*) *between the true and estimated cellular proportions under* (*A*) *different effect sizes or* (*B*) *different proportions of unknown cells. Ten random sets were generated for each condition, and the RMSE in a set was calculated with five pairs of numbers. The medians of the* 10 *random sets were connected as the line in the figure.*

then evaluate those approaches across different proportions of unknown cells. As a result, it is clear that the performance of all approaches decreases when the unknown cell content increases or the effect size decreases. The abovementioned phenomenon reflects that these gene expression deconvolution approaches are less stable in their inference for tissues with highly uncharacterized content or weak cell type-specific signal. However, the simulation shows that the effects of high unknown cell content and low effect size on BayICE estimation are less severe, and, overall, BayICE outperforms the other methods.

4.3. *Evaluating gene detection accuracy.* The spike-and-slab prior in our model provides a natural consequence of gene selection. In contrast, the existing models require an external tool to select differentially expressed genes before deconvolution, and DESeq and edgeR are two popular gene selection tools adopted by the deconvolution models. To evaluate the accuracy of gene detection, we compare with DESeq and edgeR. It is worth noting that the genes specifically expressed in the unknown cell type cannot be identified using DESeq and edgeR, and, hence, the unknown cell type-related genes are excluded in the evaluation. According to the simulation setting, each gene set is differentially expressed in one cell type. For simplicity, the effect size of marker genes is fixed at 0.2, and the cell proportions are randomly decided across samples. In this simulation study we also assume some genes express inconsistently between purified cells and the same cell type in bulk samples. For each cell type we randomly select 100 genes from the gene pool to be inconsistent genes and disturbed the expression of these inconsistent genes by multiplying an inconsistency level. The values of inconsistency level are 0.1, 0.2, 0.3, 0.4 and 0.5, and the mean of the disturbed gene expression is the original mean in Section 4.1 multiplied by 0.9, 0.8, 0.7, 0.6 or 0.5.

In this partial comparison we evaluate the accuracy of identifying genes related to the four cell types present in the reference. AUC is used to assess the gene detection under different levels of inconsistency, and the results are shown in Figure 4. For the low inconsistency

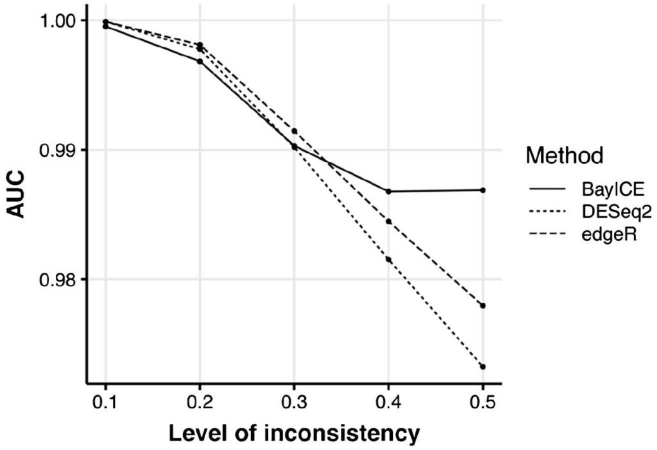

FIG. 4. *AUC for gene identification. Comparison of the AUC performance on gene identification. Measuring the performances of BayIC, DESeq2 and edgeR at identifying marker genes as measured by the Area Under the ROC Curve (AUC).*

level, the AUCs among three methods are close, and it implies that BayICE is comparable to the other approaches designed explicitly for gene selection. In the case of the considerable inconsistency in gene expression, the AUC of BayICE outperforms DESeqs2 and edgeR, and it shows that BayICE succeeds in borrowing information from mixed bulk samples for gene detection. Consequently, this partial comparison reveals that BayICE can efficiently recover the findings of the two-step approaches when cell activity causes the difference in gene expression between pure cells and tumors.

5. Validation in real data with true proportions.

5.1. *Validation in sorted cell subsets.* In Section 5.1 we consider two microarray mixture experiments for validation. They can be downloaded from the Gene Expression Omnibus database using the accession numbers GSE19830 and GSE11058 (Abbas et al. (2009); Shen-Orr et al. (2010)). The samples from GSE19830 were mixed with rat brain, liver and lung tissue derived from the same animal in different proportions. The samples in GSE11058 were mixed with four immune cell lines, Jurkat, IM-9, Raji and THP-1 at various proportions. To validate the performance in the case of the incomplete reference set, we masked one cell type from the reference set and treated the excluded type as the unknown cell component in tissues. Although EPIC is designed for RNA-seq data, the core concept of EPIC modeling is constrained least-squares optimization, and it can be widely applied to various types of data. Hence, we also performed comparisons with EPIC and NNLS on the abovementioned two real datasets.

Several microarray studies have confirmed that deconvolution on raw-scale data rather than on log-scale data can accurately reflect the underlying cellular components. The raw scale of the data was adopted in this validation study. The signature gene selection for the methods we compared in the microarray followed Hunt et al. (Hunt et al. (2018)), using a t-test between one cell type and all other cell types for each gene, and we selected the top 200 differentially expressed genes associated with each cell type.

Figures 5 and 6 compare the true and estimated cell proportions using different methods under the two evaluated datasets. For example, the scatter plot on the first row of Figure 5 represents the deconvolution results when the expression profile of Jurkat cells was unknown to all the methods. According to the characteristics of the semireference-based approach, the proportions of Jurkat cells in mixed samples can be recovered through estimating unknown

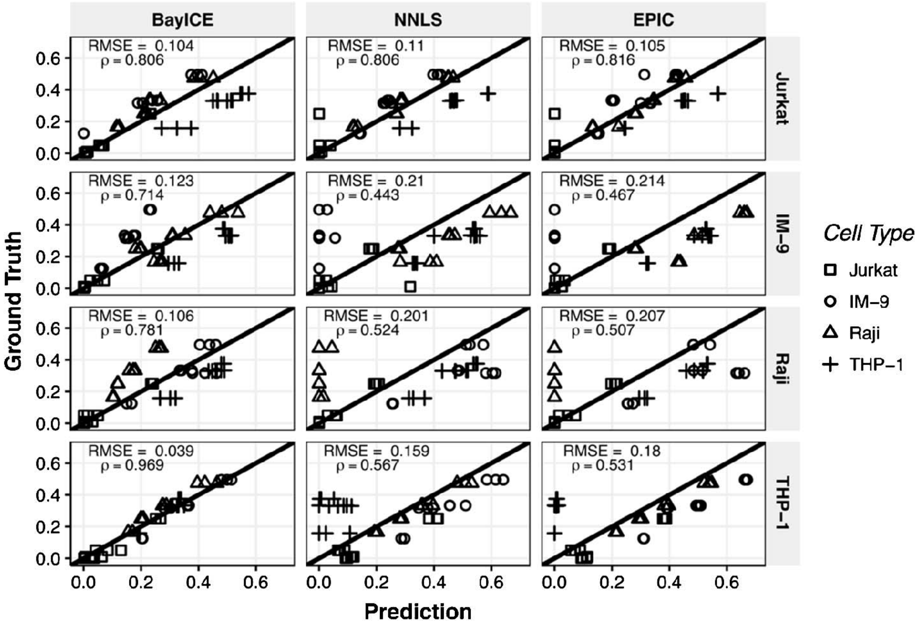

FIG. 5. *Validation by GSE*11058. *Scatter plots comparing estimated cell proportions between the true and estimated proportions from the results of GSE*11058. *Each column represents a particular method of deconvolution. The row name indicates the cell type that was masked from the reference set and referred to as the unknown cell type. Each of the* 12 *mixed sample results in four estimates of weights and thus four points in each plot. The root-mean-square error and correlation* (*ρ*) *between the ground truth and estimation are also provided in the upper-left corner of each plot.*

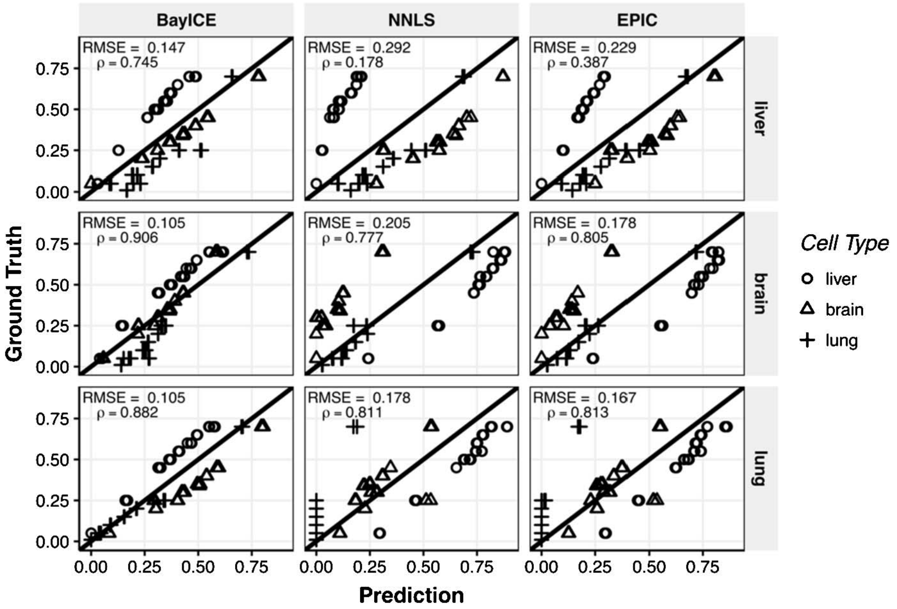

FIG. 6. *Validation by GSE*19830. *Scatter plots comparing the true and estimated proportions from the results of GSE*19830. *Each column represents a particular method of deconvolution. The row name indicates the cell type that was masked from the reference set and referred to as the unknown cell type. Each of the nine mixed sample results in three estimates of weights and thus three points in each plot. The root-mean-square error and correlation* (*ρ*) *between the ground truth and estimation are also provided in the upper-left corner of each plot.*

cell proportions. Notably, the constrained models, NNLS and EPIC, tend to assign a very small proportion to the unknown cell component. This phenomenon was observed in the Supplementary Material of robustness and can be attributed to the loss of the shift-invariant property which causes the shrinkage of unknown cell proportion estimates. In Figure 5 the estimates of BayICE for THP-1 cell type has a slight bias. The estimation bias was caused by the inconsistency of the gene expression profiles between the purified THP-1 cells and the THP-1 cells present in the mixture bulk sample. This reason can be supported by the high quality-estimation accuracy of BayICE when we treated THP-1 as the unknown cell component in tissues. Specifically, in the BayICE result in Figure 5, pink dots representing THP-1 have downward bias of predicted proportion (dots below the diagonal line) when Jurkat, IM-9 or Raji is treated as unknown, but the bias diminishes (pink dots lie on the diagonal line) when THP-1 is treated unknown.

Furthermore, we used the RMSE and Pearson correlation to assess the performance. RMSE is a measure of accuracy, and the average RMSE of BayICE is 0.093, significantly lower than those of NNLS (0.170) and EPIC (0.176). Similarly, the measurement of correlation value is used to monitor the relative order of cellular proportion estimates, and BayICE also outperformed NNLS and EPIC in terms of the relative size (BayICE = 0.82, NNLS = 0.59, and EPIC = 0.58). Apparently, BayICE possesses the advantage of incomplete data deconvolution. Furthermore, the difficulty of deconvolution increases when the abundance of the unknown cell increases. For instance, in Figure 6 the content of liver cells exceeds 50% of the mixed samples on average, and the deconvolution approaches perform relatively worse when liver cells are excluded from the reference set. The aforementioned observation is consistent with our simulation results.

5.2. *Validation in single-cell RNA sequencing.* Single-cell RNA sequencing (scRNA-seq) is a powerful technology to study the gene expression of individual cells. Many deconvolution studies have implemented strategies of constructing reference set from single-cell experiments (Newman et al. (2019); Racle et al. (2017); Wang et al. (2019)). To demonstrate the applicability of BayICE for scRNA-seq, we adopted the scRNA-seq data of melanoma from 19 melanoma samples (Tirosh et al. (2016)) which includes T cell, B cell, macrophage, cancer-associated fibroblast (CAF), endothelial cell, natural killer cell and cancer cell types. The expression profiles were measured from 4645 malignant, immune and stromal cells. The number of cells classified to each cell type in each tumor can be found in the Supplementary Material of the original paper. We randomly selected 13 out of 19 samples as the reference set and subsequently applied BayICE, EPIC and NNLS to estimate the cell proportions of the remaining six samples. The cell-specific expression profile of each sample in the reference set is reconstructed by summing the expression profiles of the same cell type. Similarly, the bulk gene expression is the sum of expression profiles of all cells in each sample. The true cell proportion of the bulk sample is proportional to the cell count of each cell type. To simulate a realistic situation of a cancer study, we masked the cancer cell-specific profiles from the reference and formed an incomplete reference set.

As a result, we obtained the estimated cell proportions by BayICE, EPIC and NNLS, and the result is illustrated in Figure 7. Accordingly, Figure 7 shows that the correlations between the true and estimated cancer cell proportions from BayICE, EPIC and NNLS are 0.949, 0.129 and 0.149, respectively. Similar to the validation study in Section 5.1, EPIC and NNLS tend to assign a very small proportion to the cancer cell component, as demonstrated in Figures S7 and S8, and, thus, these methods cannot accurately infer the cellular proportion. In addition to the cancer cell types, the result also reveals that BayICE can more accurately estimate the proportions of the other cell types than EPIC and NNLS while all methods have poor quality for CAF proportion estimation.

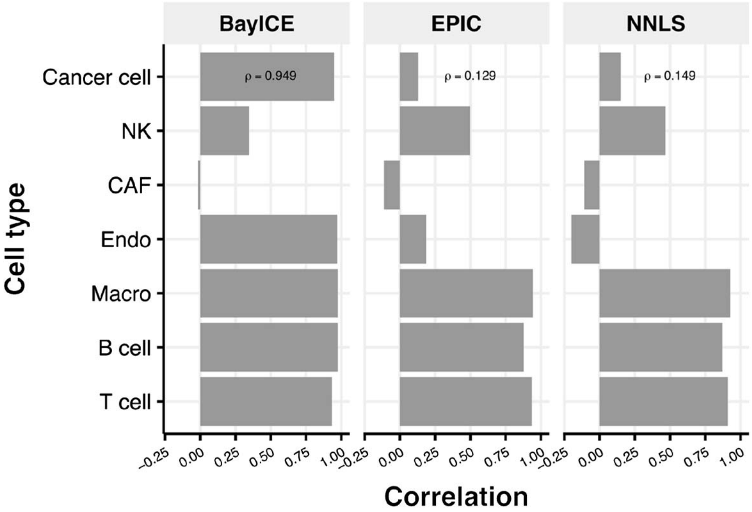

FIG. 7. *Bar plot of the correlation between the truth and estimate in single-cell RNA sequencing.*

Plot showing the correlation (ρ) between the true and estimated cell proportions for each method from single-cell RNA sequencing data of six melanoma samples. The y-axis indicates the cell types and the x-axis represents the values of the correlation. Abbreviation: Macro, macrophage; Endo, endothelial cell; CAF, cancer-associated fibroblast; NK, natural killer cell.

6. Application to nonsmall cell lung cancer. To demonstrate an application of BayICE to a biological problem, we consider RNA-seq data of lung tissues from patients with nonsmall cell lung cancer (NSCLC). Because tumor-infiltrating lymphocytes play a critical role in cancer treatment, exploring the changes in components of immune cells across tumors is of interest. Thus, we consider 199 patients with NSCLC obtained from GSE81089 and apply BayICE to estimate the cellular components of the tumor-infiltrating lymphocytes in each tumor sample (Mezheyeuski et al. (2018)). To construct the reference set, we collect RNA-seq samples of B cells, T cells, granulocytes and monocytes of blood tissues from GSE51984 (Pabst et al. (2016)). In addition to immune cell types, we include normal lung tissues in the reference set, and the expression profiles of 19 normal lung tissues from GSE81089 are used to infer the contents of normal lung cells in tumors. Because the immune cells are also present in normal lung tissues, we first apply BayICE to normal lung samples to extract the purified gene expression of lung cells and estimate the immune cell components in normal samples. Additionally, we randomly divided 19 normal samples into two sets: 10 samples are used for extracting purified expression of lung cells in Section 6.1, and nine samples are analyzed along with tumors for validation in Section 6.2. As a result, we used the complete reference set consisting of immune cell profiles and purified lung cell profiles to recover the cellular proportions of each NSCLC sample.

6.1. *Deconvolution of normal lung tissues.* Figure 8 illustrates the estimated cellular proportions of 10 normal lung samples based on the B cells, T cells, granulocytes and monocytes. The results of normal lung tissue deconvolution reveal that monocytes are more prevalent than the other immune cells in lung tissues. Monocytes typically circulate through the blood for 1–3 days before migrating into tissues, where they become macrophages or dendritic cells. In lungs, monocytes migrate from the bloodstream into the pulmonary alveolus and are specifically called alveolar macrophages which play a critical role in homeostasis, host defense, response to foreign substances and tissue remodeling (Kopf, Schneider and Nobs (2015)).

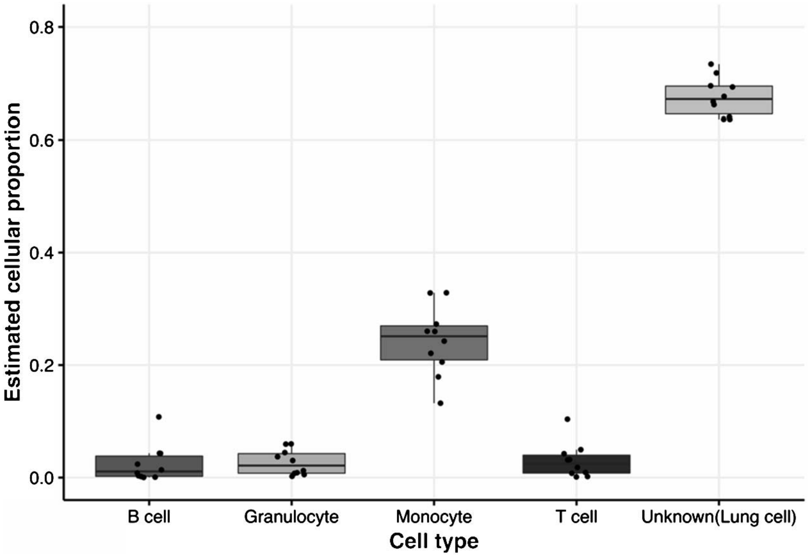

FIG. 8. *Cell proportion estimation of normal lung tissues. Plot showing estimated proportions of different cell types from* 10 *normal lung tissues. The rightmost box represents the estimated unknown component from BayICE, and we refer to this component as normal lung cells.*

In addition to the immune cell types collected in the reference set, BayICE could extract the component of one uncharacterized cell type present in tissues. In this case the unknown cell type was presumably dominated by normal lung cells, and we calculated the mean expression profile of the unknown cell type using the posterior mean. The next step was to integrate the immune cell profiles with the estimated profile of normal lung cells to construct a new reference set for tumor sample deconvolution.

6.2. *Deconvolution of tumors.* In this step we investigated the 199 NSCLC samples and nine healthy samples according to the new reference set. We applied BayICE to these 208 samples and estimated the fractions of normal lung cells, B cells, T cells, granulocytes and monocytes as well as malignant cells. Following the mechanism of normal tissue deconvolution, the malignant cell fractions were defined as the unknown cell proportions in tumor deconvolution. To more effectively understand the change in cellular components during tumor progression, we considered the classification of a malignant tumor (TNM) staging system which is a standard for classifying the extent of cancer spread. Figure 9 is a boxplot of estimated cell proportions under different TNM stages.

Two crucial observations were made from the deconvolution results. First, the estimated cell proportions were more consistent between healthy samples than between patients with NSCLC. For example, the interquartile ranges of estimated lung cell proportions in normal tissues and tumors were 0.047 and 0.137, respectively. A three-fold difference of dispersion between healthy samples and patients revealed that the homeostatic balance between cells in the lung is disturbed when tissues are cancerous. The high fluctuation of cellular components between patients with NSCLC can directly explain the intertumor heterogeneity which leads to the different attributes of different tumors despite the same diagnoses. Second, our study showed that the number of immune cells in the tissue microenvironment increases from normal stage to cancerous stage which is in agreement with past studies (Banat et al. (2015); Li et al. (2016); Ruffini et al. (2009); Seo et al. (2018)). In particular, Banat et al. (2015) comprehensively assessed the number of immune cells in lung cancer by directly counting cells with cell-specific biomarkers and observed an increased number of immune cells in lung

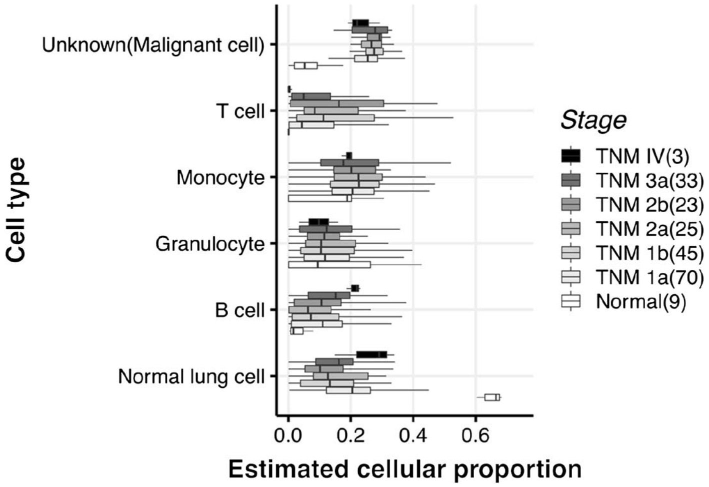

FIG. 9. *Box plot of cell proportion estimation. Orange boxes refer to the results of the nine normal lung samples, and other boxes are results from* 199 *patients with nonsmall cell lung cancer. The tumor samples can be grouped into six stages, and the sample number included for each stage is indicated in the parentheses. For samples of the same tumor stage, the estimated proportions of any specific cell type are summarized as one boxplot, and the boxplots for the same cell type are plotted side-by-side.*

cancer tissues compared with healthy donor lungs. In 2018, Seo et al. (2018) applied two approaches, ESTIMATE and TIMER, to infer the cellular components in NSCLC (Li et al. (2016); Seo et al. (2018); Yoshihara et al. (2013)). Their results showed a high abundance of dendritic cells (derived from monocytes) which coincides with our observation of estimated monocyte proportions.

7. Discussion and conclusion.

In this study, we developed a novel deconvolution model, BayICE, to infer the cellular components of bulk tissues. BayICE is a semireference-based approach that aims to explore cell populations characterized by an external reference set of purified samples and simultaneously investigate the content of uncharacterized cells present in bulk tissues. However, in contrast to constrained models, BayICE takes advantage of a hierarchical Bayesian framework to not only estimate the unknown cell proportion but also recover its gene expression profile. Furthermore, we defined the shift-invariant property for revealing the character of semireference-based deconvolution. We showed that a methodology of deconvolution maintaining the shift-invariant property, such as BayICE, can guarantee a relatively unbiased estimation for the cellular proportion. In the validation study we demonstrated that the lack of shift-invariant property essentially causes the shrinkage of the estimate for the unknown cell proportion in EPIC and NNLS.

Other than the above-mentioned merits, there are two major contributions of this study. First, BayICE integrates gene expression deconvolution and gene selection in the same model. Most of the current deconvolution approaches require preanalysis of an additional dataset to identify signature genes for deconvolution, and the external gene selection might not be consistent to the target dataset. Thus, BayICE incorporates SSVS, a Bayesian variable selection approach, to implement internal gene selection. Second, BayICE adopts shared parameters between the pure cells and tumor samples for cell-specific effect. It has been studied that the cell-to-environment interaction causes some of the genes to be expressed inconsistently after cell sorting. We have shown that the joint modeling of both pure cells and tumors in BayICE is more resistant to the problem of inconsistent genes.

To evaluate the model's performance, we first conducted an analysis under several simulation scenarios to investigate cell proportion estimation, gene identification and the robustness of the model. For proportion estimation, we compared two other semireference-based approaches, EPIC and NNLS, with BayICE under different unknown cell contents. The results revealed that BayICE significantly outperformed the other methods. We further provided a partial comparison of gene identification with the well-known gene selection approaches, DESeq and edgeR. We found that BayICE can decompose bulk data extremely well and identify cell type-related genes. Moreover, a simulation study using three different simulators revealed that BayICE is more robust with respect to data types. In addition to simulation data, we further applied two real datasets with underlying true cell proportions and one single-cell RNA-seq dataset for validation. The validation of incomplete data deconvolution was consistent with our simulation results, in which BayICE exhibited high accuracy in cell proportion estimation.

Our real data application presented an example using the data of 199 patients with NSCLC. We first applied BayICE to healthy lung samples to investigate the cellular components under a normal condition and extract the relatively purified expression profile of lung cells. The deconvolution of normal tissues succeeded in capturing the primary component of immune cells. Next, we formed a new reference set consisting of the immune cell profiles and estimated normal lung cell profiles to analyze the patients with NSCLC. From the analysis we observed inter-tumor heterogeneity in NSCLC according to the high variation in cell proportions across tumors. In addition, when comparing immune cell proportions between normal and NSCLC samples, the increased immune cell content in tumors revealed that the immune system was highly activated in the cancerous microenvironment. The inference of NSCLC deconvolution coincides with not only the analytic deconvolution results from other studies but also the observations from an immunohistochemical experiment.

BayICE has thoroughly addressed technical problems of deconvolution to investigate cellular components, but one issue relating to cell activity remains. In real applications the activities of cell-to-cell communication and cell-to-environment interaction cause some of the genes to be expressed inconsistently between reference and bulk tissue samples. This phenomenon increases the difficulty of selecting marker genes, and an inappropriate gene set limits the ability to explore tissue environments. Although the joint modeling technique of BayICE can adjust the biased gene expression induced by cell sorting, integrating the biological information of cell activity with deconvolution is believed to be more efficient in estimating cell proportions of bulk tissues. Single-cell RNA-seq has emerged as a powerful new set of technologies for characterizing cell interaction, and the primary goal of our future studies is incorporating single-cell RNA data to modify the prior structure with cell interaction. This extension will generate new insights into the deconvolution framework.

Moreover, we will also investigate the unknown cell proportions estimated by BayICE. The proportion of the unknown cell class can be further decomposed if the unknown class is comprised of multiple cancerous cell types. Instead of estimating immune cell proportions, the study of tumor clonal evolution focuses on exploring the contents of different tumor cell types to understand tumor progression. Hence, after BayICE dissects the part belonged to tumor, we can further decompose tumor cells to evaluate the size of tumor clones.

All of the work in this study was implemented on R, and the R package, BayICE, is publicly available for deconvolution analysis (https://github.com/AshTai/BayICE). The computing time to perform deconvolution of 199 NSCLC samples is about 45 minutes with 16 CPU cores, 2.3 GHz and 32 GB RAM.

Acknowledgments. We thank Dr. Jiebiao Wang (University of Pittsburgh) for the suggestions to refine the manuscript. This work was supported by the Ministry of Science and Technology [MOST 107-2118-M-007-001].

SUPPLEMENTARY MATERIAL

Supplementary material (DOI: 10.1214/20-AOAS1376SUPP; .pdf). We provide a comprehensive discussion of the shift-invariant property, the procedure of data normalization, the detail of the Gibbs sampler in BayICE, the two remaining simulators, the convergence property of BayICE, and the robustness for deconvolution.

REFERENCES

ABBAS, A. R., WOLSLEGEL, K., SESHASAYEE, D., MODRUSAN, Z. and CLARK, H. F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS ONE* **4** e6098. https://doi.org/10.1371/journal.pone.0006098

ANGHEL, C. V., QUON, G., HAIDER, S., NGUYEN, F., DESHWAR, A. G., MORRIS, Q. D. and BOUTROS, P. C. (2015). ISOpureR: An R implementation of a computational purification algorithm of mixed tumour profiles. *BMC Bioinform.* **16** 156.

BANAT, G., TRETYN, A., PULLAMSETTI, S. S., WILHELM, J., WEIGERT, A., OLESCH, C., EBEL, K., STIEWE, T., GRIMMINGER, F. et al. (2015). Immune and inflammatory cell composition of human lung cancer stroma. *PLoS ONE* **10** e0139073. https://doi.org/10.1371/journal.pone.0139073

BECHT, E., GIRALDO, N. A., LACROIX, L., BUTTARD, B., ELAROUCI, N., PETITPREZ, F., SELVES, J., LAURENT-PUIG, P., SAUTÈS-FRIDMAN, C. et al. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17** 218.

CHAROENTONG, P., FINOTELLO, F., ANGELOVA, M., MAYER, C., EFREMOVA, M., RIEDER, D., HACKL, H. and TRAJANOSKI, Z. (2017). Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.* **18** 248–262. https://doi.org/10.1016/j.celrep.2016.12.019

CHEN, Z., HUANG, A., SUN, J., JIANG, T., QIN, F. X.-F. and WU, A. (2017). Inference of immune cell composition on the expression profiles of mouse tissue. *Sci. Rep.* **7** 40508.

CHEN, Z., QUAN, L., HUANG, A., ZHAO, Q., YUAN, Y., YUAN, X., SHEN, Q., SHANG, J., BEN, Y. et al. (2018). seq-ImmuCC: Cell-centric view of tissue transcriptome measuring cellular compositions of immune microenvironment from mouse RNA-Seq data. *Front. Immunol.* **9** 1286. https://doi.org/10.3389/fimmu.2018.01286

DIEU-NOSJEAN, M.-C., GOC, J., GIRALDO, N. A., SAUTÈS-FRIDMAN, C. and FRIDMAN, W. H. (2014). Tertiary lymphoid structures in cancer and beyond. *Trends Immunol.* **35** 571–580.

FRIDMAN, W. H., PAGÈS, F., SAUTÈS-FRIDMAN, C. and GALON, J. (2012). The immune contexture in human tumours: Impact on clinical outcome. *Nat. Rev. Cancer* **12** nrc3245.

FRIDMAN, W. H., ZITVOGEL, L., SAUTÈS–FRIDMAN, C. and KROEMER, G. (2017). The immune contexture in cancer prognosis and treatment. *Nat. Rev. Clin. Oncol.* **14** 717.

GAUJOUX, R. and SEOIGHE, C. (2012). Semi-supervised nonnegative matrix factorization for gene expression deconvolution: A case study. *Infect. Genet. Evol.* **12** 913–921.

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GONG, T., HARTMANN, N., KOHANE, I. S., BRINKMANN, V., STAEDTLER, F., LETZKUS, M., BONGIOVANNI, S. and SZUSTAKOWSKI, J. D. (2011). Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS ONE* **6** e27156. https://doi.org/10.1371/journal.pone.0027156

HU, P., ZHANG, W., XIN, H. and DENG, G. (2016). Single cell isolation and analysis. *Front. Cell Dev. Biol.* **4** 116. https://doi.org/10.3389/fcell.2016.00116

HUNT, G. J., FREYTAG, S., BAHLO, M. and GAGNON-BARTSCH, J. A. (2018). dtangle: Accurate and robust cell type deconvolution. *Bioinformatics* **35** 2093–2099.

ISHWARAN, H., KOGALUR, U. B. and RAO, J. S. (2010). spikeslab: Prediction and variable selection using spike and slab regression. *R J.* **2**.

ISHWARAN, H. and RAO, J. S. (2005a). Spike and slab gene selection for multigroup microarray data. *J. Amer. Statist. Assoc.* **100** 764–780. MR2201009 https://doi.org/10.1198/016214505000000051

ISHWARAN, H. and RAO, J. S. (2005b). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158 https://doi.org/10.1214/009053604000001147

KOPF, M., SCHNEIDER, C. and NOBS, S. P. (2015). The development and function of lung-resident macrophages and dendritic cells. *Nat. Immunol.* **16** 36–44. https://doi.org/10.1038/ni.3052

LI, B., SEVERSON, E., PIGNON, J.-C., ZHAO, H., LI, T., NOVAK, J., JIANG, P., SHEN, H., ASTER, J. C. et al. (2016). Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. *Genome Biol.* **17** 174. https://doi.org/10.1186/s13059-016-1028-7

MEZHEYEUSKI, A., BERGSLAND, C. H., BACKMAN, M., DJUREINOVIC, D., SJÖBLOM, T., BRUUN, J. and MICKE, P. (2018). Multispectral imaging for quantitative and compartment-specific immune infiltrates reveals distinct immune profiles that classify lung cancer patients. *J. Pathol.* **244** 421–431. https://doi.org/10.1002/path.5026

NEWMAN, A. M., LIU, C. L., GREEN, M. R., GENTLES, A. J., FENG, W., XU, Y., HOANG, C. D., DIEHN, M. and ALIZADEH, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12** 453–457. https://doi.org/10.1038/nmeth.3337

NEWMAN, A. M., STEEN, C. B., LIU, C. L., GENTLES, A. J., CHAUDHURI, A. A., SCHERER, F., KHODADOUST, M. S., ESFAHANI, M. S., LUCA, B. A. et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37** 773–782. https://doi.org/10.1038/s41587-019-0114-2

OGUNDIJO, O. E. and WANG, X. (2017). A sequential Monte Carlo approach to gene expression deconvolution. *PLoS ONE* **12** e0186167.

PABST, C., BERGERON, A., LAVALLÉE, V.-P., YEH, J., GENDRON, P., NORDDAHL, G. L., KROSL, J., BOIVIN, I., DENEAULT, E. et al. (2016). GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo. Blood:blood-2015-2011-683649.

PRASSAS, I., CHRYSTOJA, C. C., MAKAWITA, S. and DIAMANDIS, E. P. (2012). Bioinformatic identification of proteins with tissue-specific expression for biomarker discovery. *BMC Med.* **10** 39.

RACLE, J., DE JONGE, K., BAUMGAERTNER, P., SPEISER, D. E. and GFELLER, D. (2017). Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6**. https://doi.org/10.7554/eLife.26476

RICHARDSON, G. M., LANNIGAN, J. and MACARA, I. G. (2015). Does FACS perturb gene expression? *Cytometry, Part A* **87** 166–175.

RUFFINI, E., ASIOLI, S., FILOSSO, P. L., LYBERIS, P., BRUNA, M. C., MACRÌ, L., DANIELE, L. and OLIARO, A. (2009). Clinical significance of tumor-infiltrating lymphocytes in lung neoplasms. *Ann. Thorac. Surg.* **87** 365–372.

SEO, J.-S., KIM, A., SHIN, J.-Y. and KIM, Y. T. (2018). Comprehensive analysis of the tumor immune microenvironment in non-small cell lung cancer for efficacy of checkpoint inhibitor. *Sci. Rep.* **8** 14576.

SHEN-ORR, S. S., TIBSHIRANI, R., KHATRI, P., BODIAN, D. L., STAEDTLER, F., PERRY, N. M., HASTIE, T., SARWAL, M. M., DAVIS, M. M. et al. (2010). Cell type-specific gene expression differences in complex tissues. *Nat. Methods* **7** 287–289. https://doi.org/10.1038/nmeth.1439

SUN, X., SUN, S. and YANG, S. (2019). An efficient and flexible method for deconvoluting bulk RNA-seq data with single-cell RNA-seq data. *Cells* **8** 1161.

TAI, A.-S., TSENG, G. C. and HSIEH, W.-P. (2021). Supplement to "BayICE: A Bayesian hierarchical model for semireference-based deconvolution of bulk transcriptomic data." https://doi.org/10.1214/20-AOAS1376SUPP

TIROSH, I., IZAR, B., PRAKADAN, S. M., WADSWORTH, M. H., TREACY, D., TROMBETTA, J. J., ROTEM, A., RODMAN, C., LIAN, C. et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352** 189–196.

TRAN, H. T. N., ANG, K. S., CHEVRIER, M., ZHANG, X., LEE, N. Y. S., GOH, M. and CHEN, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21** 1–32.

VAN DEN BRINK, S. C., SAGE, F., VÉRTESY, Á., SPANJAARD, B., PETERSON-MADURO, J., BARON, C. S., ROBIN, C. and VAN OUDENAARDEN, A. (2017). Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14** 935–936. https://doi.org/10.1038/nmeth.4437

WAGNER, G. P., KIN, K. and LYNCH, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131** 281–285.

WANG, X., PARK, J., SUSZTAK, K., ZHANG, N. R. and LI, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10** 1–9.

WHITTEMORE, A. S. (2007). A Bayesian false discovery rate for multiple testing. *J. Appl. Stat.* **34** 1–9. MR2345755 https://doi.org/10.1080/02664760600994745

XIE, F., ZHOU, M. and XU, Y. (2018). BayCount: A Bayesian decomposition method for inferring tumor heterogeneity using RNA-Seq counts. *Ann. Appl. Stat.* **12** 1605–1627. MR3852690 https://doi.org/10.1214/17-AOAS1123

YANG, Y. H., DUDOIT, S., LUU, P., LIN, D. M., PENG, V., NGAI, J. and SPEED, T. P. (2002). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30** e15–e15.

YOSHIHARA, K., SHAHMORADGOLI, M., MARTÍNEZ, E., VEGESNA, R., KIM, H., TORRES-GARCIA, W., TREVIÑO, V., SHEN, H., LAIRD, P. W. et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4** 2612. https://doi.org/10.1038/ncomms3612

ZHONG, Y., WAN, Y.-W., PANG, K., CHOW, L. M. and LIU, Z. (2013). Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinform.* **14** 89.