HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2021 January 01.

Published in final edited form as:

J Am Stat Assoc. 2020 ; 115(531): 1055–1065. doi:10.1080/01621459.2019.1654874.

ICeD-T Provides Accurate Estimates of Immune Cell Abundance in Tumor Samples by Allowing for Aberrant Gene Expression Patterns

Douglas R. Wilson[a], **Chong Jin**[a], **Joseph G. Ibrahim**[a], **Wei Sun**[a,b,c,d]

[a]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC

[b]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA

[c]Department of Biostatistics, University of Washington, Seattle, WA

Abstract

Immunotherapies have attracted lots of research interests recently. The need to understand the underlying mechanisms of immunotherapies and to develop precision immunotherapy regimens has spurred great interest in characterizing immune cell composition within the tumor microenvironment. Several methods have been developed to estimate immune cell composition using gene expression data from bulk tumor samples. However, these methods are not flexible enough to handle aberrant patterns of gene expression data, e.g., inconsistent cell type-specific gene expression between purified reference samples and tumor samples. We propose a novel statistical method for expression deconvolution called ICeD-T (Immune Cell Deconvolution in Tumor tissues). ICeD-T automatically identifies aberrant genes whose expression are inconsistent with the deconvolution model and down-weights their contributions to cell type abundance estimates. We evaluated the performance of ICeD-T versus existing methods in simulation studies and several real data analyses. ICeD-T displayed comparable or superior performance to these competing methods. Applying these methods to assess the relationship between immunotherapy response and immune cell composition, ICeD-T is able to identify significant associations that are missed by its competitors.

Keywords

Immunotherapy; Immuno-Oncology; Bulk Expression; Deconvolution; RNA-seq; Microarray

1 Introduction

The evolving relationship between a cancer and its host's immune system is well summarized by a hypothesis known as immunoediting. Immunoediting stresses that the immune system not only suppresses tumor cells, but also shapes tumor immunogenicity in ways that may promote tumor growth [1, 2]. For example, consider the relationship between tumors and tumor-infiltrating T cells. Infiltrating T cells can be cytotoxic, contributing to the

[d]Contact: wsun@fredhutch.org.

reduction of cancer cell populations. However, these T cells also express immune checkpoints that inhibit their function, and such checkpoints can prevent the immune system from indiscriminately attacking healthy host cells. Under selective pressure from the immune system, cancers cells may exploit the immune checkpoints to escape the attack by infiltrating T cells.

Immunotherapy were developed based on the insights of immunoediting [3]. Among the best-known immunotherapy strategies, immune checkpoint inhibitors block immune inhibition pathways that restrict effective anti-tumor T cell responses [4]. Checkpoint inhibitors have achieved phenomenal success in a fraction of cancer patients, exhibiting response rates around 40% and 20% for melanoma and lung cancer, respectively [5]. It is of great clinical interest to identify the subset of cancer patients who may respond to checkpoint inhibitors. Use of tumor-infiltrating immune cells to predict clinical response to therapy has shown promising results. Previous studies have shown that the patients with CD8+ T cells around tumor cells have higher response rate to checkpoint inhibitors [6]. In addition to benefiting the development of precision immunotherapies, immune cell composition in tumor samples have also demonstrated prognostic value [7, 8]. Therefore, studying immune cell composition in tumor samples is timely and potentially has high impact on cancer research.

Several groups have studied immune cell composition using gene expression data from bulk tumor samples [9–14]. These pioneering works have demonstrated promising results, but also bear some limitations. For example, a subset of these works estimate immune cell presence using the expression of few genes [9, 10], or calculate average expression of the genes that are highly expressed in certain cell type [15] instead of estimating immune cell composition. As an alternative, several methods have been proposed to estimate immune cell composition using a regression-based approach, with gene expression from bulk tumor samples as the response variable and reference gene expression from purified cell types as covariates. For example, CIBERSORT [12] employs support-vector regression. TIMER [13] uses a linear regression and removes the genes with very high expression due to their strong influence on model fitting. EPIC [14] uses weighted linear regression to give the genes with lower expression variation higher weights. These regression-based methods, when applied to tumor expression data, explicitly or implicitly assume that they start with a set of genes that have negligible expression in tumor cells, and that the expression of immune cells is conserved between purified reference samples and tumor samples. These assumptions are questionable as many factors that affect gene expression may differ between tumor and reference samples.

In this paper, we propose a new statistical method for cell type deconvolution entitled ICeD-T, which stands for Immune Cell Deconvolution in Tumor tissues. ICeD-T is an extension of existing regression-based methods [12–14] with two major novel features designed to overcome the limitations of these methods. First, ICeD-T employs a likelihood-based framework, which assumes that gene expression follows a log-normal distribution. An earlier work has shown that deconvolution should be performed on linear-scale instead of log-scale of gene expression data since linear-scale mixing of gene expression better captures the biological reality [16]. However, since gene expression variation increases with

expression level, genes with higher expression may become outliers with great influence on linear scale deconvolution. Therefore, one may need to remove genes with high expression for robust deconvolution analysis [13]. The log transformation, often used in expression studies, enjoys variance-stabilizing and skew-mitigation properties that limit the impact of genes with high expression [17, 18]. ICeD-T is able to perform gene expression deconvolution on the linear-scale while simultaneously incorporating the beneficial properties of the log-transformation through our method design and the use of log-normal distribution. CIBESORT also performs gene expression deconvolution on linear scale, and its epsilon-insensitive L1 loss function helps limit the impact of genes with high expression. In contrast, EPIC uses an L2 loss function and may be more sensitive to the variation of genes with high expression. See Supplementary Materials Section A.1 for a more details of CIBERSORT and EPIC.

Second, ICeD-T automatically identifies the genes whose expression in tumor samples are inconsistent with reference profiles (referred to as aberrant genes) and down-weights the contribution of such aberrant genes in cell type abundance estimation. The aberrant expression of those genes may be due to altered expression in tumor infiltrating immune cells or unexpected tumor cell expression. Since CIBESORT uses epsilon-insensitive L1 loss function, genes with loss smaller than a threshold (epsilon) do not contribute to cell type composition estimation. However, this gene-selection property of CIBESORT is very different from aberrant gene detection of ICeD-T. CIBESORT ignores the genes that fit the model very well. In contrast, ICeD-T down-weights those genes that fit the model poorly.

2 Statistical Methods

2.1 The Input Data

ICeD-T can be applied on both microarray data and RNA-seq data. The gene expression from bulk tumor samples and purified samples of each cell type should be normalized in a consistent manner. For example, quantile normalization can be applied for microarray data. For RNA-seq data, we may use FPKM (Fragments Per Kilobase of transcript per Million mapped reads), FPKM-UQ, or TPM (Transcript per Million). More specifically, to calculate FPKM, we divide gene expression (# of RNA-seq fragments) by total number of mapped fragments per sample (in millions) and the gene length (in kilo bases). FPKM-UQ is a variant of FPKM where sample-specific read-depth is measured by 75 percentile of gene-level fragment counts across all genes, instead of the total number of mapped fragments. TPM reverses the order of the two normalization steps. It first divides the gene-level fragment counts by gene length, and then divides it by the summation of gene-length corrected fragment counts across all genes.

Additional information utilized by ICeD-T's deconvolution model includes a pre-selected gene set (ideally, genes with immune-specific expression) and tumor purity, if available. Several such gene sets have been prepared by previous works, such as the gene sets used by CIBERSORT or EPIC [12, 14]. Provision of tumor purity is optional, and it can be computed, for example, using somatic copy number alteration data [19].

2.2 Statistical Model

We first define some notations. We use Y and Z to denote the observed gene expression data from bulk tumor samples and purified reference samples, respectively. X denotes the unobserved cell type-specific expression in bulk tumor samples. Let n be the number of bulk samples, J be the number of genes, and K be the number of cell types. Y is a matrix of size $n \times J$, where Y_{ij} is the expression of gene j in the i-th bulk tumor sample. Z is a three-dimensional array, and Z_{jkh} is the expression of gene j in the h-th purified sample of cell type k plus a small constant, such as 1/6, so that $Z_{jkh} > 0$. $1 \leq h \leq H_k$, where H_k is the number of purified samples of cell type k.

Specification of the ICeD-T model begins with a consideration of expression behavior in purified references samples of constituent cell types. We assume that the Z_{jkh}'s follow independent log-normal distributions, given by:

$$\log(Z_{jkh}) \sim \mathcal{N}(\mu_{jk}, \sigma^2_{jk}), \tag{1}$$

and

$$E[Z_{jkh}] = \gamma_{jk} = \exp(\mu_{jk} + \sigma^2_{jk}/2), \text{ and } V[Z_{jkh}] = \gamma^2_{jk}[\exp(\sigma^2_{jk}) - 1]. \tag{2}$$

Therefore, the distribution parameters for each cell type's gene expression (e.g., μ_{jk} and σ^2_{jk}) may be estimated by the mean and variance of the log-transformed Z_{jkh} values. Once estimated, these parameters represent expression profiles for each cell type in our deconvolution model. Optionally, ICeD-T accepts previously computed profiles that would replace the γ_{jk} above.

Shift focus to the n bulk tumor samples. Assuming that each sample is composed of K non-tumor cell types, with K specified *a prior*. We assume Y_{ij}, the expression of gene j in bulk tumor sample i, can be modeled as weighted summation of gene expression in constituent cell types with a multiplicative error term:

$$Y_{ij} = \epsilon_{ij} \sum_{k=1}^{K} \rho_{ik} X_{ijk}, \tag{3}$$

where ρ_{ik} is the proportion of expression attributable to cell type k and ϵ_{ij} follows a log-normal distribution with mean value of $\log(\epsilon_{ij})$ being 0. If tumor purity information is provided, $\sum_{k=1}^{K} \rho_{ik} = 1 - \rho_{iT}$, where ρ_{iT} is tumor purity. If tumor purity is not provided, $\sum_{k=1}^{K} \rho_{ik} \leq 1$.

We begin to develop the probabilistic framework utilized by ICeD-T to model the relationship posited above by first assuming that there are no aberrant genes (i.e., gene expression of each cell type in reference samples is consistent with gene expression in tumor microenvironment). Under such an assumption, X_{ijk} has the same distribution as the Z_{jkh} for any i, h, and j (i.e., $X_{ijk} \sim Z_{jkh}$) summation of independent log-normal random variables does

not have a closed form distribution function. To address this issue, ICeD-T approximates the distribution of Y_{ij} using another log-normal:

$$\log(Y_{ij}) \sim \mathcal{N}\left(\tilde{\mu}_{ijC}, \Delta_j \sigma^2_{iC}\right), \text{ where } \tilde{\mu}_{ijC} = \log\left(\sum_{k=1}^{K} \rho_{ik}\gamma_{jk}\right) - \Delta_j \sigma^2_{iC}, \quad (4)$$

and Δj is the weight for the j-th gene.

The approximation used above is based upon the Fenton-Wilkinson approach which states that the summation of log-normal's can be approximated by another log-normal whose parameters are obtained via moment-matching [20]. Under a strict Fenton-Wilkinson approach, the distribution of Y_{ij} would be given by:

$$\log(Y_{ij}) \sim \mathcal{N}\left(\tilde{\mu}_{ijC}, \tilde{\sigma}^2_{ijC}\right)$$

where

$$\tilde{\mu}_{ijC} = \log\left(\sum_{k=1}^{K} \rho_{ik}\gamma_{jk}\right) - \tilde{\sigma}^2_{ijC}/2,$$
$$\tilde{\sigma}^2_{ijC} = \log\left(\sum_{k=1}^{K} (\rho_{ik}\gamma_{jk})^2\left[\exp\left(\sigma^2_{jk}\right) - 1\right] \Big/ \left[\sum_{k=1}^{K} \rho_{ik}\gamma_{jk}\right]^2 + 1\right).$$

We replace the variance structure posited by Fenton-Wilkinson with the weighted variance model of equation (4) as the weighted model demonstrates improved fit and stability in simulated data.

Regarding the variance weights used by ICeD-T, we implement two different options. One assumes a homogeneous weight for all genes, i.e., $\Delta_j = 1$ for all j. Later we refer to this option as "no weight". For the other option of "with weight", we define weights to be proportional to maximal cell-type-specific variances. Specifically, let σ_j^{*2} be the maximum expression variance of gene j across all cell types:

$$\sigma_j^{*2} = \max_k\left(\hat{\sigma}^2_{jk}\right).$$

The weight of a gene j is then specified as follows:

$$\Delta_j = \frac{\sigma_j^{*2}}{\text{median}_j\left[\sigma_j^{*2}\right]}.$$

Thus, a gene's weight compares its maximal expression variance to the median of all such maxima across genes. Under this construction, genes with larger variances will have larger

variance weights. Larger variance weights ensure that residuals from such genes will have smaller impact on estimation of cell type composition.

The Δj specified above requires slight modification to improve stability of the model fit. Unadjusted, this procedure can provide some genes with excessively small variance weights and some genes with excessively high variance weights. To control this extreme behavior, the bottom 15% of variance weights are replaced with the 15th percentile variance weight across all genes. Similarly, the top 15% of all variance weights are replaced by the 85th percentile variance weight. In this way, no genes are allowed to become too minimally or maximally important to model fit.

Return to the specification of Y_{ij} in equation (4). Now assume that some genes in the dataset are aberrant. The expression of these aberrant genes is inconsistent with the deconvolution model. For aberrant genes, ICeD-T borrows the expression structure proposed for consistent genes but inflates the variance. Thus, if gene j is aberrant, the expression of Y_{ij} is given by:

$$\log(Y_{ij}) \sim \mathcal{N}\left(\tilde{\mu}_{ijA}, \Delta_j \sigma^2_{iA}\right), \tag{5}$$

where

$$\tilde{\mu}_{ijA} = \log\left(\sum_{k=1}^{K} \rho_{ik}\gamma_{jk}\right) - \Delta_j \sigma^2_{iA} \text{ and } \sigma^2_{iA} > \sigma^2_{iC}.$$

By allowing aberrant genes to have larger variance, the ICeD-T model flattens the likelihood for such genes, and thus down-weights their contributions to cell type proportion estimates.

Direct use of the likelihoods provided by equations (4) and (5) within bulk data is impossible since it is unknown whether a gene is consistent or aberrant *a priori*. Thus, ICeD-T must model expression at any gene as a mixture of the log-normal distributions pertaining to consistent and aberrant genes. The mixture likelihood utilized by ICeD-T is found below:

$$Y_{ij} \sim p_i \mathcal{LN}\left(\tilde{\mu}_{ijC}, \Delta_j \sigma^2_{iC}\right) + (1 - p_i)\mathcal{LN}\left(\tilde{\mu}_{ijA}, \Delta_j \sigma^2_{iA}\right), \tag{6}$$

where $\mathcal{LN}$ denotes the density function of a log-normal distribution, and pi and $1–p_i$ denotes the proportion of genes being consistent and aberrant, respectively. The density function of observed data y_{ij} is

$$f(Y_{ij}) = \frac{p_i}{y_{ij}\sigma_{iC}\sqrt{2\pi\Delta_j}} \exp\left\{-\frac{\left[\log(y_{ij}) - \tilde{\mu}_{ijC}\right]^2}{2\Delta_j \sigma^2_{iC}}\right\} + \frac{1 - p_i}{y_{ij}\sigma_{iA}\sqrt{2\pi\Delta_j}} \exp\left\{-\frac{\left[\log(y_{ij}) - \tilde{\mu}_{ijA}\right]^2}{2\Delta_j \sigma^2_{iA}}\right\}$$

The constant terms $1/\left(y_{ij}\sqrt{2\pi}\right)$ can be omitted in MLE.

This likelihood function can be maximized using an EM algorithm. Missing data are introduced in the form of class membership indicators H_{ij}, where $H_{ij} = 0$ or 1 denotes

whether the j-gene is aberrant or consistent in the i-th bulk tumor sample, respectively. Thus, the complete data log-likelihood for the i-th bulk tumor sample is given by:

$$\ell_i = \sum_{j=1}^{J} H_{ij}\left[\log(p_i) - (1/2)\log\left(\Delta_j\sigma_{iC}^2\right) - \left(1/2\Delta_j\sigma_{(C)}^2\right)\left(\log\left(y_{ij}\right) - \tilde{\mu}_{jC}\right)^2\right] + \left(1 - H_{ij}\right)\left[\log(1-p_i) - (1/2)\log\left(\Delta_j\sigma_{iA}^2\right) - \left(1/2\Delta_j\sigma_{iA}^2\right)\left(\log\left(y_{ij}\right) - \tilde{\mu}_{ijA}\right)^2\right],$$

where J is the number of genes used in our model.

There are altogether $K + 3$ parameters to estimate for each bulk tumor sample, including K cell type compositions $(\rho_{i1},\ldots,\rho_{iK})$, two variance parameters σ_{iC}^2 and σ_{iA}^2, and the mixture proportion p_i. The sample size is the number of genes J. K is usually smaller than 20, and J can be a few hundreds. Therefore, we do not expect identifiability issues as long as there is no strong co-linearity in cell type-specific expression data. There are constraints on the values of ρ_{ik}'s such that $\rho_{ik} \geq 0$ and $\sum_k \rho_{ik} \leq 1$, and we incorporate these constraints when maximizing the log-likelihood function.

Before fitting the mixture model, we perform an initial model fitting to estimate cell type composition and calculate the residuals for each gene after accounting for cell type composition. We initialize the aberrant group of genes by those with larger residuals from initial model fitting. When the EM algorithm converges, we may redefine the aberrant vs. consistent group based on the final estimates of the variance terms. In our experience, the aberrant vs. consistent assignments rarely switch.

Within each EM step, maximization of Q function with respect to $\left(\rho_{i1},\ldots,\rho_{iK},\sigma_{iC}^2,\sigma_{iA}^2\right)$ and p_i are separable. Given the other parameters, the estimate of p_i has a closed form. Given p_i, the remaining parameters are grouped into two blocks: the mixture proportions ρ_{ik}'s (block 1) and the two variance parameters $\left(\sigma_{iC}^2,\sigma_{iA}^2\right)$ (block 2). The parameters of these two blocks are iteratively updated. Given the estimates of $\left(\sigma_{iC}^2,\sigma_{iA}^2\right)$ the mixture proportions ρ_{ik} are estimated using numerical optimization (the BFGS algorithm) while the constraints are incorporated using the Augmented Lagrangian method (R function auglag). Given the estimates of the mixture proportions ρ_{ik}'s, the two variance terms $\left(\sigma_{iC}^2,\sigma_{iA}^2\right)$ are involved in separate pieces of the complete data log-likelihood, and thus can be estimated separately. Given variance weights, each of σ_{iC}^2 and σ_{iA}^2 is estimated by numerical optimization (R function `optimize`). Without variance weights, they can be estimated by closed form. See Supplementary Materials Section A for details of the parameter estimation steps.

The ρ_{ik}'s estimated by any regression-based deconvolution approach should be interpreted as the proportion of gene expression contributed by certain cell types. If one seeks to estimate the proportion of cells, these ρ_{ik}'s should be adjusted by cell size factors. We borrow the cell size factors, denoted by s_k, from Racle et al. [14] and construct revised relative abundance of immune cell types by $\rho_{ik}^* = (\rho_{ik}/s_k)/\sum_{i=1}^{K}(\rho_{ik}/s_k)$. Further details are provided in the Supplementary Materials (Section C.2).

3 Results

3.1 Simulation Study

We conducted a simulation study to evaluate the performance of ICeD-T, CIBERSORT, and EPIC. For each method, we seek to assess the estimation accuracy and the robustness of estimation in the presence of aberrant gene behavior. For ICeD-T only, we also assess its ability to identify aberrant genes.

We simulated reference expression of 250 genes for 5 cell types: one tumor cell type and four immune cell types. Our simulations assume that these 250 genes were selected to be expressed in immune cells but not tumor cells. When there are no aberrant genes, the expression of these 250 genes in a bulk tumor sample was simulated by mixing the 4 immune cell types with known proportions. For each gene, we assume it is expressed in one of the four immune cell types and has low/background expression in the other three immune cell types. To better mimic the complexity of real data, we do not assume one homogeneous background expression. Instead, we assume the background expression has a three-tiered scale to reflect lowly, moderately or highly expressed genes (range in log scale: 2.0–8.0). Average log-transformed expression for the expressed cell type was simulated by an up-shift of background expression level (range: 3.5–9.0). See Supplementary Materials Section B.1 for more details. Using RNA-seq expression data from immune cells taken from Linsley et al. [21], a mean-variance relationship was computed from FPKM-UQ normalized data across immune specific genes. Then we used this mean-variance relationship and the simulated average expression profiles to decide corresponding variance with allowance for random error. Fifteen reference samples were simulated for each cell type from its unique expression profile using a log-normal distribution.

To generate the expression of a bulk tumor sample, a tumor purity value was simulated from a normal distribution (mean=0.60, sd=0.15) and truncated at endpoints of 0.17 and 0.95. The remaining immune cell proportions were then simulated from a Dirichlet distribution with average abundances ranging from 15% to 40%. For each gene in the bulk tumor sample, its expression in each immune cell type was simulated from a log-normal distribution and a weighted summation of these expression values was computed as the expression in the bulk tumor sample. For the simulation setup with aberrant behavior of gene expression, approximately twenty percent of genes were randomly selected as aberrant genes. Among them, 25% had down-regulated expression in the highly expressed cell type, 25% had up-regulated expression of the highly expressed cell type, and 50% had expression in tumor cells at a background level. See the Appendix B for further details regarding the construction of these simulations and additional simulation results.

The expression profile of each cell type was estimated from the 15 simulated samples of that cell type. This reference is used for deconvolution in each of the following models: ICeD-T without variance weights, ICeD-T with variance weights, LNorm with variance weights, CIBERSORT (version 1.06), and EPIC. LNorm stands for “log normal”, and it is a variant of the ICeD-T model which does not consider aberrant gene behavior.

When there is no aberrance in gene expression, all methods perform well, while ICeD-T provides the most accurate estimates of cell type proportions (Figure 1). When 20% of the 250 genes are aberrant, the performance of LNorm, EPIC, and CIBERSORT all become worse, while the performance of ICeD-T method remain similar (Figure 2). Both EPIC and LNorm's cell type proportion estimates suffer from bias and larger variance in the presence of aberrant genes. CIBERSORT still performs relatively well, but has an apparent inflation of the estimation variance. While the weighted version of ICeD-T provides the best results, both weighted and unweighted ICeD-T are able to maintain high accuracy with minimal estimation variance (Figure 2(a)–(b)).

To identify aberrant genes, ICeD-T computes the posterior probability of a gene being consistent. Examining the distribution of this quantity across consistent and aberrant genes, we see that both the weighted and unweighted versions of ICeD-T separate consistent and aberrant genes reasonably well (Figure 3). The weighted version of ICeD-T provides more accurate estimate of the proportion of aberrant genes, and identifies consistent genes with higher confidence. For aberrant genes, the posterior probabilities of being consistent show a bi-modal distribution, implying that a small proportion of aberrant genes are missed. This is partly due to our very challenging simulation setting, with three types of aberrant patterns and three tiers of expression levels for background genes, which diminishes the difference between background cell types and expressed cell types.

We have also conducted simulation studies to evaluate the robustness of our method when giving different initial values of cell type compositions. Our final estimates of cell type compositions are virtually the same across 1,000 initial values (Supplementary Figure 8), suggesting the likelihood surface is concave or very close to be concave.

3.2 Validation in Microarray Expression of PBMCs

In the CIBERSORT paper, Newman et al. [12] described the collection of peripheral blood mononuclear cell (PBMC) gene expression data from 20 healthy adults. After extraction of PBMC samples from each subject, these samples were subjected to microarray expression analysis and flow cytometric measurement to establish ground-truth of cell type proportions. We use this dataset to evaluate our method and compare its performance with CIBERSORT and EPIC.

To be consistent with the approach used by Newman et al. [12], we use their LM22 reference of cell type-specific gene expression for all methods. The LM22 reference matrix is derived from microarray gene expression data, and thus is consistent with the gene expression platform of the bulk tissue samples. The reference matrices of EPIC (TRef/BRef for tumor samples and normal samples, respectively) were derived from RNA-seq data, and thus they are inappropriate in microarray settings. Since EPIC and ICeD-T expect the gene expression of bulk samples and reference samples are measured on the same scale, we normalized gene expression data from bulk samples by quantile normalization to match the expression data used to derive the LM22 matrix. The results of each method are then restricted to the nine cell-types examined in Newman et al. [12]: naive B-cells, memory B-cells, CD8+ T-cells, naive/memory resting/memory activated CD4+ T-cells, $\gamma\delta$ T-cells, Natural killer cells, and monocytes. Estimates for each mixture sample are renormalized so

that their summation equals 100 after correction for cell sizes of different cell types. The accuracy of each method is assessed by comparing sums of squared errors and correlations between the expression-based cell type proportion estimates and flow-cytometry estimates. Correlations are computed by pooling cell type proportions for all subjects and all cell types.

Examining the results of the 9 original cell types, ICeD-T provides the most accurate estimates of cell type proportions in terms of sum of squared errors. CIBERSORT, on the other hand, provides the most accurate estimates with respect to the correlations (Table 1, Figure 4). However, the superior correlation of CIBERSORT is due in part to several cell subsets with positive correlations but severe bias (e.g., memory activated CD4 T-cells, memory resting CD4 T-cells) (Supplementary Materials Section C.4). After grouping a few highly similar cell types (e.g., grouping naive B-cells and memory B-cells as B cells, and naive/memory resting/memory activated CD4+ T-cells as CD4+ T cells), ICeD-T achieves comparable or higher correlation between expression-based cell type proportion estimates and flow-cytometry estimates while maintaining the smallest sum of squared errors (Table 1, Figure 4). In this dataset, EPIC has very poor performance, which may be due to the fact that it is designed for RNA-seq data.

3.3 Flow Cytometry Validation in Melanomas

In the EPIC paper, Racle et al. [14] obtained metastatic melanoma samples from the lymph nodes of four patients with stage III melanomas. A portion of each of these samples was used for a flow cytometric analysis while the remaining portion was used for bulk RNA-sequencing. Results from flow cytometry were used to establish a ground-truth cell type composition. TPM-normalized RNA-seq expressions and flow cytometry measured compositions were extracted directly from the EPIC R package.

We used EPIC's TRef matrix as reference gene expression for both EPIC and ICeD-T. ICeD-T was run in four different modes, with or without variance weights (denoted by wY and wN, respectively) and with or without sample purity as part of the inputs (denoted by pY and pN, respectively). For this analysis, purity is defined as the proportion of non-immune content plus the proportions of cells not assessed via flow cytometry (e.g. macrophages, fibroblasts, and endothelia's, and others). CIBERSORT was fit using both the LM22 and TRef matrices directly to the TPM data. All cell type proportion estimates were corrected by cell size factors reported by Racle et al. [14]. To allow comparison of ICeD-T and EPIC with CIBERSORT that only computes relative immune cell abundance estimates, we obtain relative proportions for all methods by normalizing cell type proportions so that they add up to 1.

Overall EPIC provides more accurate estimates of the absolute proportions of all immune cells, while ICeD-T provides more accurate estimation of the relative proportions of immune cells among the modeled immune cell types (Table 2, Figure 5, Supplementary Materials, Section D.3). Comparing relative proportions of immune cells, ICeD-T (pY, wY) improves upon EPIC's fit in terms of the overall sum of squared error (0.043 vs 0.11) while preserving strong correlation (0.924 vs 0.918) across all subjects.

We also evaluated the performance of CIBERSORT versus the flow cytometry estimates. Compared with other methods, CIBERSORT has comparable or less accurate estimates of cell type proportions in three subjects, but much better performance than the other methods in subject LAU125 (Table 2). Based on flow cytometry estimates, this subject has somewhat unexpected immune cell proportion: almost entirely B-cells. All methods perform much worse in this subject than other subjects, with larger sum squared errors. CIBERSORT's relatively better performance for this challenging subject could be due to a combination of its objective function and use of LM22 reference matrix. CIBERSORT's performance becomes worse when using TRef instead of LM22 as reference matrix, though it still has much smaller sum squared error than EPIC and ICeD-T.

3.4 Application to anti-PD-1 Immunotherapy Data

Finally, we use ICeD-T, CIBERSORT, and EPIC to analyze an RNA-seq dataset from bulk tumor samples of melanoma patients [22]. The RNA-seq data are available in 28 patients before treatment with pembrolizumab. We seek to associate treatment response (Complete Response, Partial Response, or Non-response) using CD8+ cell type composition estimated by each of the three methods.

Fastq files of RNA-seq data were downloaded from NCBI Sequence Read Archive, mapped to human genome (hg38) and the number of RNA-seq fragments per gene were counted. Then such counts were normalized by TPM. We ran EPIC and ICeD-T using the TRef reference gene expression data. ICeD-T was fit without using tumor purity as this information was not available. CIBERSORT was fit using LM22 reference matrix. Abundance estimates across each method are corrected using EPIC's cell type size factors. In addition, to ensure comparability across all methods, immune cell proportions are renormalized so that their summation equals to 1.

Differences in relative CD8+ T-cell abundance across response categories was assessed using a Jonckheere-Terpstra test for trended differences. The Jonckheere-Terpstra test can be considered as an extension of non-parametric ANOVA tests (e.g. Kruskal-Wallis) to allow greater power to detect ordered population differences [23]. Previous studies have shown that those cancer patients with more CD8+ T cells within tumor microenvironment are more likely to respond to anti-PD-1 treatment [24]. Thus, as one moves across response categories from most to least responsive to therapy, one would expect to see a decrease in CD8+ T cell abundance.

CIBERSORT and EPIC capture the expected relationship between CD8+ T cell proportion and immunotherapy response to some extent, but have trouble in separating the members of at least two groups. For CIBERSORT, individuals in the partial response group behave similarly to those in the progressive disease group. For EPIC, individuals in the complete response group behave similarly to those who exhibited partial response. The Jonckheere-Terpstra tests provide numerical confirmation of these difficulties as the tests are not significant, with p-values for CIBERSORT and EPIC being 0.30 and 0.14, respectively.

ICeD-T, on the other hand, provides clear visual distinction between these three groups, and shows less CD8+ T cells for those who do not respond to anti-PD-1 treatment. This

relationship is reinforced through the significant Jonckheere-Terpstra test (p=0.038). Introduction of variance weights further separates these categories (p=0.017). Cell type proportions estimates by either version of ICeD-T have higher within group similarities than either CIBERSORT or EPIC.

We also examined the probability being consistent for each gene in each sample. It appears that in this dataset it is challenging to clearly assign a gene in a sample to be consistent or aberrant with high confidence (Supplementary Figure 12). However, the relative scale of probability being consistent is still very informative. For example, our deconvolution model fits the data very well (correlation 0.87 or 0.92 when using weight or not) when the probability being consistent is high (top 1/3 of observations). In contrast, when the probability being consistent is low (bottom 1/3 of observations), the correlations between model fit and observed gene expression drop to 0.64, either using weight or not (Supplementary Figure 13).

4 Discussion

In this paper, we have outlined ICeD-T, a novel statistical method for cell type proportion estimation using gene expression from tumor tissues. ICeD-T utilizes the variance stabilizing properties of the log-transformation while simultaneously controlling for aberrant gene behavior within the tumor tissue. In addition, ICeD-T incorporates a variance weighting structure which diminishes the impact of highly variable genes on abundance estimation. Optionally, ICeD-T can refine cell type abundance estimation through use of tumor purity information, if available.

We have demonstrated that ICeD-T is an accurate model in both simulated and real datasets. The robustness of ICeD-T to misbehaved genes and its ability to identify these genes was demonstrated in simulated data. ICeD-T's accuracy was reinforced in real datasets using both microarray and RNA-seq expression where it was consistently a top performer compared with other methods. We applied ICeD-T to study the relation between CD8+ T cell proportion and response to anti-PD-1 immunotherapy and found significant associations between CD8+ T cell proportions and patients' response to immunotherapy.

ICeD-T uses a mixture model to divide all the genes into two groups: consistent ones and aberrant ones. This mixture model allows us to down-weight the contributions of aberrant genes for cell type composition estimation. The same cell type composition information is shared across all the genes, and those aberrant genes can still be used to estimate cell type composition. Therefore, this mixture model does not introduce identifiability issues. However, it is important to monitor the proportion of genes classified as aberrant, and a high proportion of aberrant genes may imply inappropriate input of cell type-specific expression.

There is room to improve the performance of ICeD-T. One direction is to refine the reference matrix of cell type-specific gene expression. In this paper, we have adopted the reference gene expression matrix (TRef) used by EPIC. TRef was constructed using single cell RNA-seq (scRNA-seq) data from melanoma cancer samples. Cell type-specific expression was estimated by pooling cells of the same cell types, identified by clustering

method. However, some technical limitations of scRNA-seq, such as dropout (expression of many genes are measured at 0 while they may be lowly expressed) may lead to biased gene expression estimates [25]. Careful examination of such effects may improve the reference matrix of cell type-specific gene expression. It may also worth considering re-estimating cell type-specific gene expression and in such cases, a hierarchical model may be considered to borrow information across genes to improve robustness.

Another future direction to improve ICeD-T is to refine the weight for each gene. We have implemented the weight for each gene based on the maximum of cell type-specific variances. Other options that use the variances across all cell types may be more desirable. However, with limited cell type-specific gene expression data, we have not yet identified a clear choice.

The ICeD-T methodology has been implemented in an R software package and it is available at https://github.com/Sun-lab/ICeDT.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

[1]. Dunn GP, Koebel CM, and Schreiber RD (2006) Interferons, immunity and cancer immunoediting. Nature Reviews Immunology, 6(11), 836.

[2]. Schreiber RD, Old LJ, and Smyth MJ (2011) Cancer immunoediting: integrating immunity?s roles in cancer suppression and promotion. Science, 331(6024), 1565–1570. [PubMed: 21436444]

[3]. Farkona S, Diamandis EP, and Blasutig IM (2016) Cancer immunotherapy: the beginning of the end of cancer?. BMC medicine, 14(1), 73. [PubMed: 27151159]

[4]. Sharma P and Allison JP (2015) The future of immune checkpoint therapy. Science, 348(6230), 56–61. [PubMed: 25838373]

[5]. Yarchoan M, Hopkins A, and Jaffee EM (2017) Tumor mutational burden and response rate to PD-1 inhibition. New England Journal of Medicine, 377(25), 2500–2501. [PubMed: 29262275]

[6]. Topalian SL, Taube JM, Anders RA, and Pardoll DM (2016) Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. Nature Reviews Cancer, 16(5), 275. [PubMed: 27079802]

[7]. Sato E, Olson SH, Ahn J, Bundy B, Nishikawa H, Qian F, Jungbluth AA, Frosina D, Gnjatic S, Ambrosone C, et al. (2005) Intraepithelial CD8+ tumor-infiltrating lymphocytes and a high CD8+/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer. Proceedings of the National Academy of Sciences of the United States of America, 102(51), 18538–18543. [PubMed: 16344461]

[8]. Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf AC, Angell H, Fredriksen T, Lafontaine L, Berger A, et al. (2013) Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity, 39(4), 782–795. [PubMed: 24138885]

[9]. Brown SD, Warren RL, Gibb EA, Martin SD, Spinelli JJ, Nelson BH, and Holt RA (2014) Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. Genome research, 24(5), 743–750. [PubMed: 24782321]

[10]. Rooney MS, Shukla SA, Wu CJ, Getz G, and Hacohen N (2015) Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell, 160(1), 48–61. [PubMed: 25594174]

[11]. Angelova M, Charoentong P, Hackl H, Fischer ML, Snajder R, Krogsdam AM, Waldner MJ, Bindea G, Mlecnik B, Galon J, et al. (2015) Characterization of the immunophenotypes and

antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. Genome biology, 16(1), 64. [PubMed: 25853550]

[12]. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, and Alizadeh AA (2015) Robust enumeration of cell subsets from tissue expression profiles. Nature methods, 12(5), 453–457. [PubMed: 25822800]

[13]. Li B, Severson E, Pignon J-C, Zhao H, Li T, Novak J, Jiang P, Shen H, Aster JC, Rodig S, et al. (2016) Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biology, 17(1), 174. [PubMed: 27549193]

[14]. Racle J, deJonge K, Baumgaertner P, Speiser DE, and Gfeller D (2017) Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife, 6, e26476. [PubMed: 29130882]

[15]. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, Selves J, Laurent-Puig P, Sautès-Fridman C, Fridman WH, et al. (2016) Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome biology, 17(1), 218. [PubMed: 27765066]

[16]. Zhong Y and Liu Z (2012) Gene expression deconvolution in linear space. Nature Methods, 9(1), 8–9.

[17]. Law CW, Chen Y, Shi W, and Smyth GK (2014) voom: Precision weights unlock linear model analysis tools for rna-seq read counts. Genome biology, 15(2), R29. [PubMed: 24485249]

[18]. Gierliński M, Cole C, Schofield P, Schurch NJ, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson G, Owen-Hughes T, et al. (2015) Statistical models for rna-seq data derived from a two-condition 48-replicate experiment. Bioinformatics, 31(22), 3625–3630. [PubMed: 26206307]

[19]. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, et al. (2012) Absolute quantification of somatic dna alterations in human cancer. Nature biotechnology, 30(5), 413–421.

[20]. Fenton L (1960) The sum of log-normal probability distributions in scatter transmission systems. IRE Transactions on Communications Systems, 8(1), 57–67.

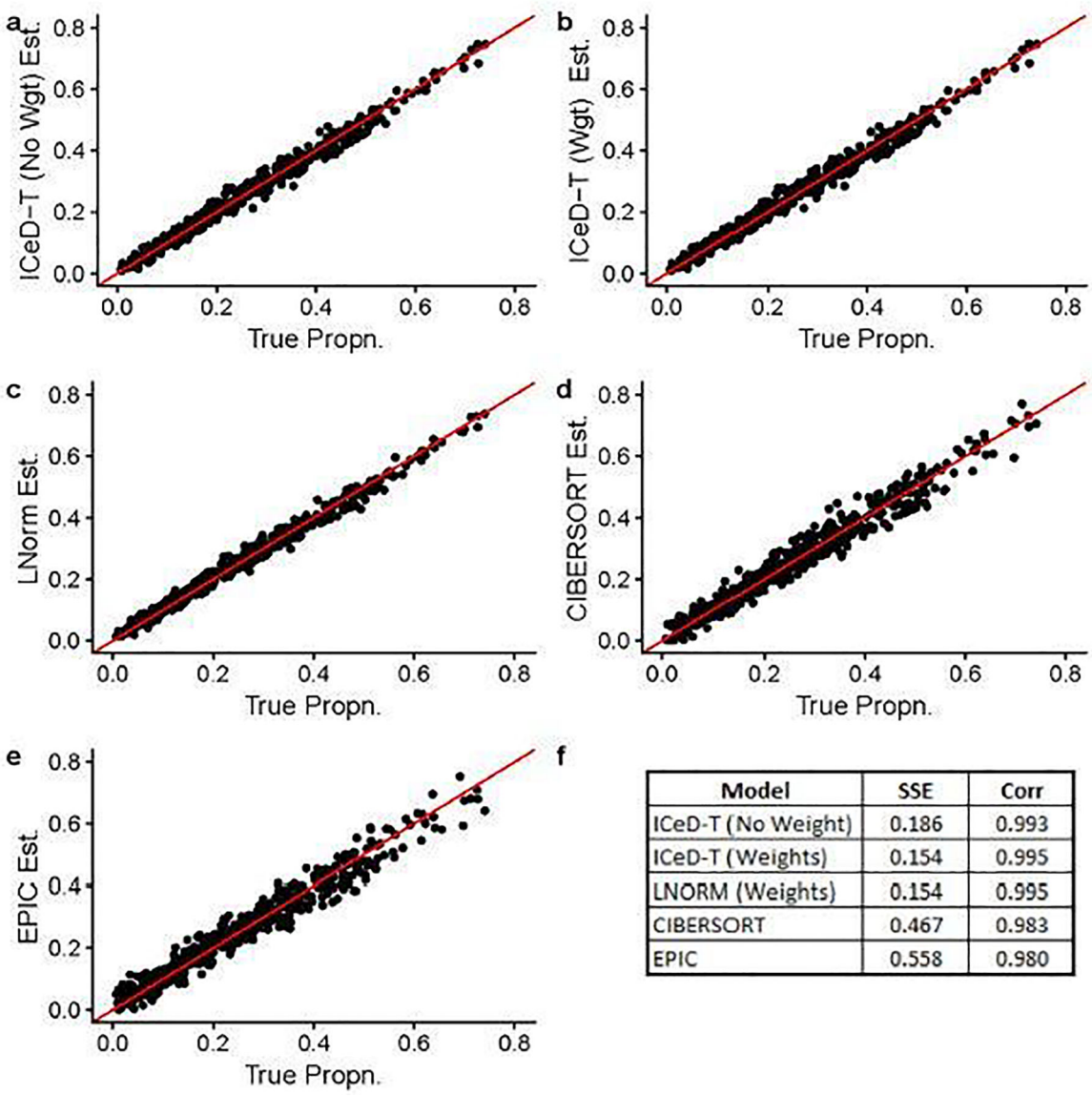
[21]. Linsley PS, Speake C, Whalen E, and Chaussabel D (2014) Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis. PloS one, 9(10), e109760. [PubMed: 25314013]

[22]. Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S, Berent-Maoz B, Pang J, Chmielowski B, Cherry G, et al. (2016) Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. Cell, 165(1), 35–44. [PubMed: 26997480]

[23]. Bewick V, Cheek L, and Ball J (4, 2004) Statistics review 10: Further nonparametric methods. Critical Care, 8(3), 196–199. [PubMed: 15153238]

[24]. Chen DS and Mellman I (2017) Elements of cancer immunity and the cancer–immune set point. Nature, 541(7637), 321. [PubMed: 28102259]

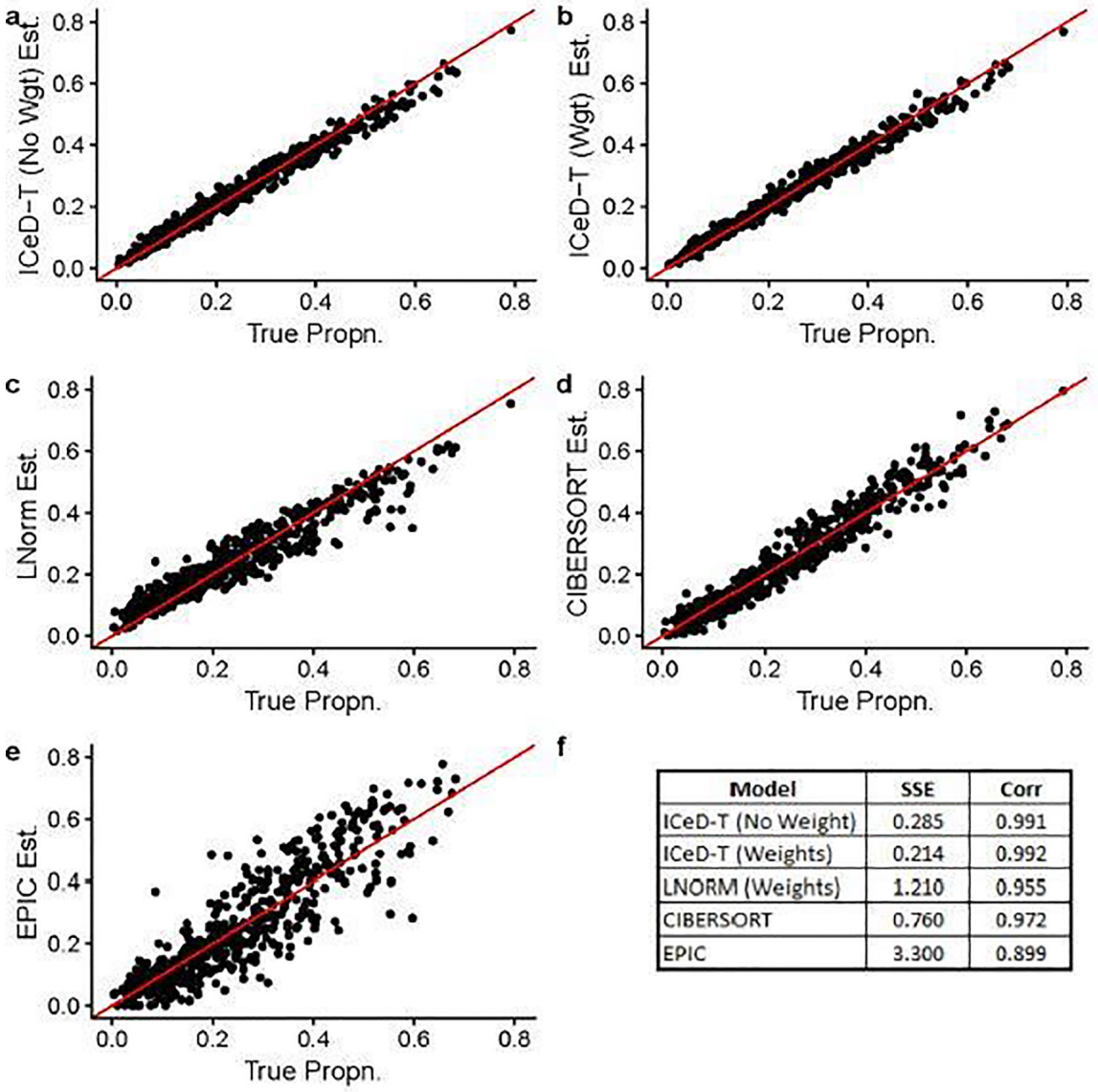
[25]. Angerer P, Simon L, Tritschler S, Wolf FA, Fischer D, and Theis FJ (2017) Single cells make big data: New challenges and opportunities in transcriptomics. Current Opinion in Systems Biology, 4, 85–91.



Model	SSE	Corr
ICeD-T (No Weight)	0.186	0.993
ICeD-T (Weights)	0.154	0.995
LNORM (Weights)	0.154	0.995
CIBERSORT	0.467	0.983
EPIC	0.558	0.980

Fig. 1.

Visualized results of model fits on simulated data without aberrance. Each dot is an estimate of ρ_{ik} (the proportion of cells attributable to cell type k, for sample i), where $i = 1,\ldots,135$ and $k = 1,\ldots,4$. Figure (f) summarizes the accuracy across all 135 samples for each model.



Model	SSE	Corr
ICeD-T (No Weight)	0.285	0.991
ICeD-T (Weights)	0.214	0.992
LNORM (Weights)	1.210	0.955
CIBERSORT	0.760	0.972
EPIC	3.300	0.899

Fig. 2.

Visualized results of model fits on simulated data when ~ 20% of the genes are aberrant. Each dot is an estimate of ρ_{ik} (the proportion of cells attributable to cell type k, for sample i), where $i = 1,\ldots,135$ and $k = 1,\ldots,4$.Figure (f) summarizes the accuracy across all 135 samples for each model.

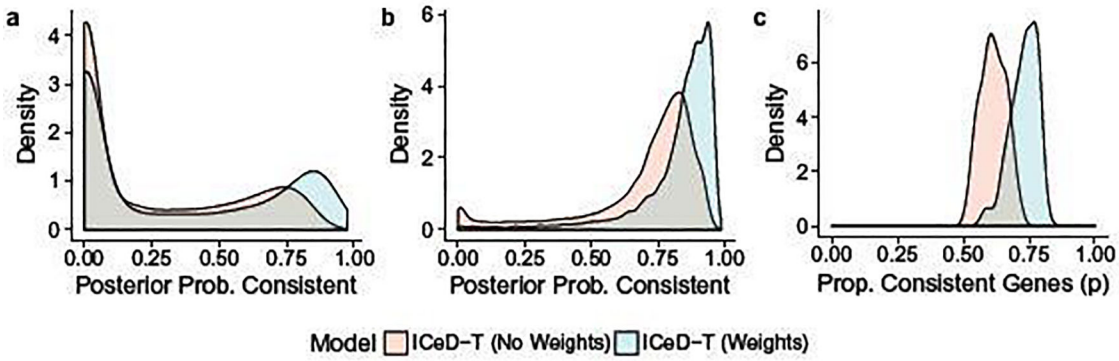


Fig. 3.

(a) The posterior probabilities of being consistent for those aberrant genes. (b) The posterior probabilities of being consistent for those consistent genes. (c) Estimates of the proportion of consistent genes.

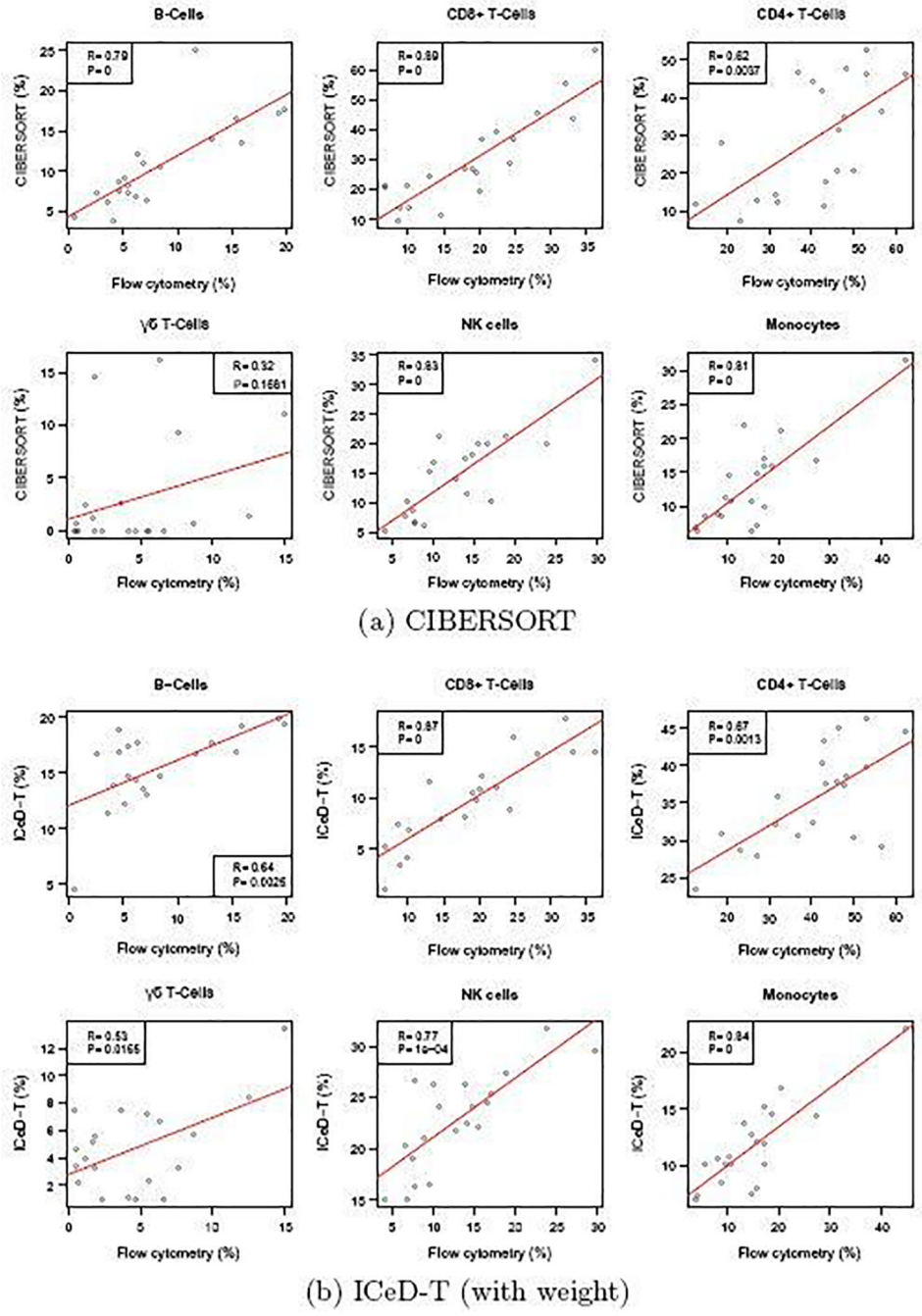


Fig. 4.

Comparison of cell type proportion estimates by CIBERSORT and ICeD-T versus the cell type proportions measured by flow cytometry. Each dot is an estimate of the relative proportion of the specified cell type in one of the 20 samples. Red lines indicate the least squares model fit to the estimated immune proportions.

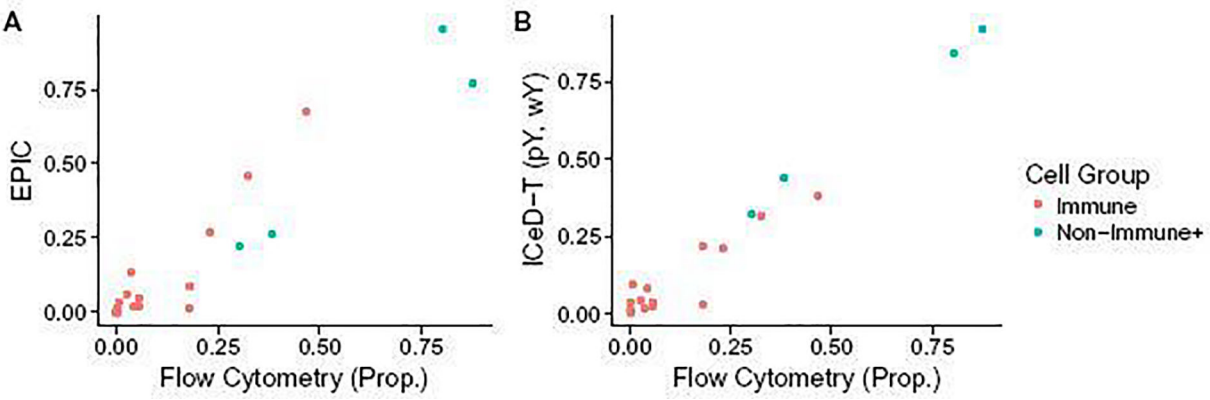


Fig. 5.

Plots of EPIC and ICeD-T model estimates against flow cytometry estimates. ICeD-T is fit using variance weights and sample purity. Each dot is an estimate of relative cell type proportions for four immune cell types (B cells, CD4+ T cells, CD8+ T cells, and NK cells) and the other types of cells in four individuals.

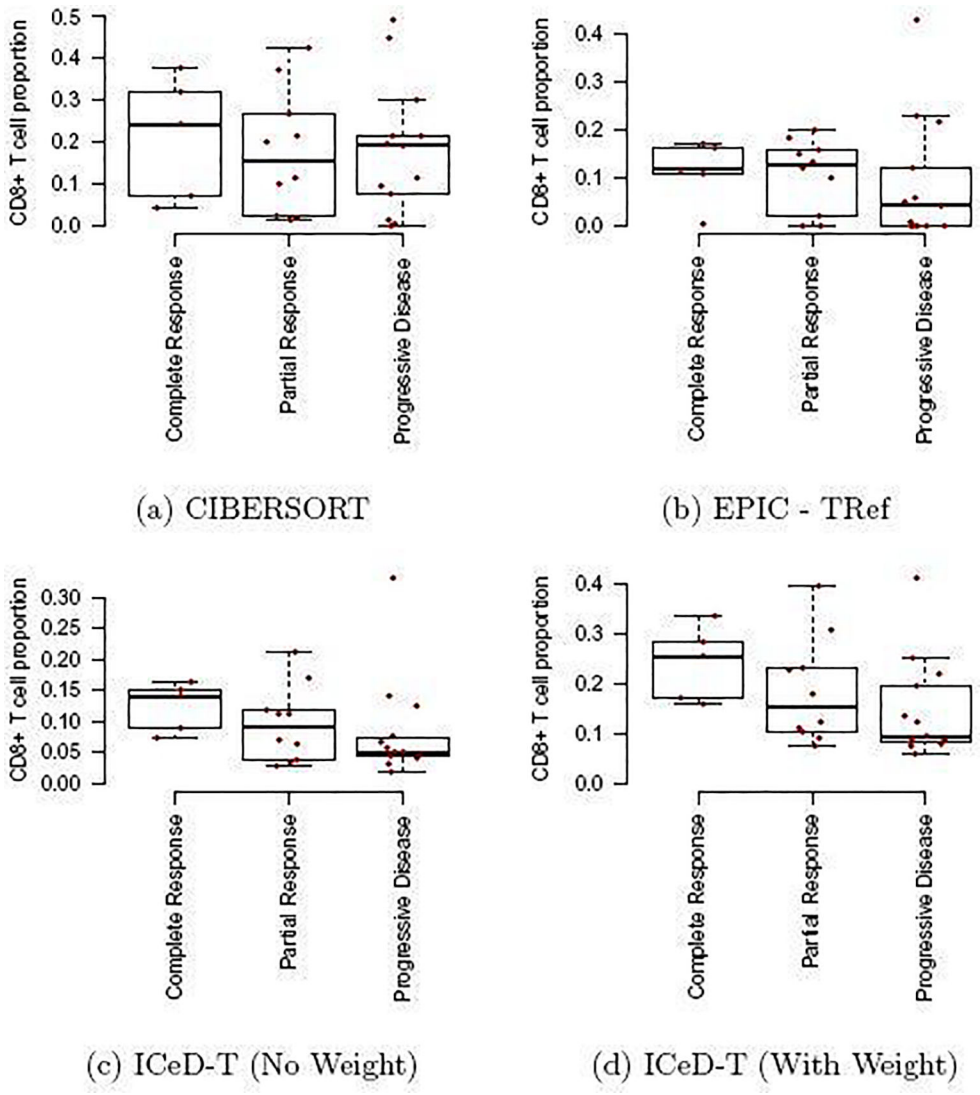


Fig. 6.

Comparison of model fits to PD-1 Immunotherapy Data. The x-axis are three response groups. The y-axis is the proportion of CD8+ T cells (i.e., parameters ρ_{ik} in our model) estimated by different methods.

Table 1

Validation of immune cell proportion estimates by flow cytometry for 9 cell types [left] and 6 cell types after grouping naive B-cells and memory B-cells as B cells, and naive/memory resting/memory activated CD4+ T-cells as CD4+ T cells [right]. SSE stands for sum squared error, and Cor stands for correlation.

Model	**SSE**	**Cor**
ICeD-T (no weight)	13.10	0.53
ICeD-T (w/ weight)	12.05	0.59
CIBERSORT	14.15	0.65
EPIC	29.43	0.31
Model	**SSE**	**Cor**
ICeD-T (no weight)	10.48	0.75
ICeD-T (w/ weight)	9.44	0.78
CIBERSORT	11.02	0.77
EPIC	32.01	0.18

Table 2

Sum of Squared Errors for relative immune proportions among all immune cell types. ICeD-T fits are labeled with (pX, wX) to indicate use of purity (pY=Yes and pN=No) and weight (wY=Yes and wN=No).

Model	LAU125	LAU1255	LAU1314	LAU335
CIBERSORT (LM22)	0.12	0.16	0.003	0.010
CIBERSORT (TRef)	0.32	0.10	0.021	0.095
EPIC	0.86	0.15	0.066	0.013
ICeD-T (pN, wN)	1.03	0.10	0.042	0.003
ICeD-T (pN, wY)	1.07	0.14	0.005	0.004
ICeD-T (pY, wN)	0.85	0.08	0.039	0.008
ICeD-T (pY, wY)	0.85	0.14	0.020	0.002