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

Contents

A Supplementary Methods

A.1 Overview of CIBERSORT and EPIC

CIBERSORT [1] employs nu-support vector regression (ν-SVR) to estimate cell type composition. Let y_j and z_{jk} be the observed expression of gene j in a bulk sample, and the k-th cell type, respectively. Then an intuitive (albeit not rigorous) definition of the loss function of a support vector regression is

$$\sum_{j \in J_\epsilon} |y_j - \sum_k z_{jk}\beta_k| + \lambda \sum_k \beta_k^2, \tag{1}$$

where λ is a tuning parameter for the penalty of regression coefficients, and J_ϵ denotes those genes such that $|y_j - \sum_k z_{jk}\beta_k| > \epsilon$, i.e., the so-called support vectors. The loss function is referred to as ϵ insensitive loss function because those genes with loss smaller than ϵ does not contribute to model fit. It is difficult to select an appropriate ϵ since it requires the information on the sizes of losses across all genes, and such information is often unavailable. An alternative choice is to reformulate the loss function by ν-SVR, which asymptotically choses ν proportion of genes as support vectors. In the implementation of CIBERSORT, three values were considered for ν: $\nu = \{0.25, 0.5, 0.75\}$. The one that leads to best model fit is chosen. This model setup does not guarantee all the regression coefficients to be non-negative. CIBERSORT sets all the negative regression coefficients to 0, and normalize the remaining ones to sum to 1. These zero-truncated and normalized regression coefficients are their final estimates of relative cell type fractions.

Using the same notations of y_j and z_{jk} for consistency, the loss function of EPIC [2] is a constrained L2 loss:

$$\sum_j w_j (y_j - \sum_k z_{jk}\beta_k)^2, \quad \beta_k \geq 0, \ \sum_{k=1}^K \beta_k \leq 1. \tag{2}$$

where w_j is the weight for the j-th gene. $w_j = \min(u_j, 100\text{median}(u_j))$, $u_j = \sum_{k=1}^K z_{jk}/(v_{jk} + \epsilon)$ for a small number ϵ to avoid division by 0, and v_{jk} the variance of the expression of gene j in the k-th cell type. In other words, genes with higher expression (after standardizing gene expression by its variance) have higher weights, and an upper bound of weights is set at $100\text{median}(u_j)$. EPIC allows an uncharacterized cell type with proportion $\beta_{K+1} = 1 - \sum_{k=1}^K \beta_k$. Finally, the estimate of β_k is the proportion of expression from cell type k rather than the proportion of cells. To obtain the proportion of cells, β_k is normalized by a cell size factor.

A.2 Notations and overview of ICeD-T

A.2.1 Notation Table

The following table contains the notation used to develop and mathematically interrogate the ICeD-T model and its variants. Subscripts for aberrant genes, denoted by $(\cdot)$ in the following table, may take values (A) or (C); (A) indexes quantities pertaining to aberrant genes and (C) indexes those in consistent genes.

Model Design Quantities		
Value	**Dimension**	**Description**
n	1×1	The number of mixed cell type samples for deconvolution.
n_k	1×1	The number of purified samples of cell type k.
K	1×1	The number of constituent cell types, excluding the tumor.
n_G	1×1	The number of signature genes used in cell type modeling.
Pure Sample Quantities		
Value	**Dimension**	**Description**
μ_{jk}	1×1	Mean log-transformed expression of gene j in cell type k.
σ^2_{jk}	1×1	The variance of log-transformed expression of gene j in cell type k.
γ_{jk}	1×1	The mean expression of gene j in cell type k on the untransformed scale.
$\boldsymbol{\gamma}$	$n_G \times K$	Matrix of mean expression across all genes and cell types.
$\boldsymbol{\gamma}_j$	$K \times 1$	Vector of mean expressions of gene j across the K cell types (j-th row of $\boldsymbol{\gamma}$).
Z_{jkh}	1×1	Normalized expression of gene j in purified sample h of cell type k.
$\mathbf{Z}_k$	$n_G \times n_q$	Collection of Z_{jkh} across all genes and purified samples.
Mixture Sample Quantities		
Value	**Dimension**	**Description**
$\tilde{\mu}_{ij(\cdot)}$	1×1	Mean expression of gene j in mixture sample i
ρ_{ik}	1×1	Proportion of RNA expression attributable to cells of type k in mixture i.
ρ_i	$K \times 1$	Collection of ρ_{ik} across cell types for subject i only.
$\sigma^2_{ij(\cdot)}$	1×1	Variance of expression for gene j in mixture sample i.
$\sigma^2_{i(\cdot)}$	1×1	Unweighted variance parameter governing expression in mixture sample i.
Δ_j	1×1	Optional variance weight for gene j.
Y_{ij}	1×1	Normalized expression of gene j in mixture sample i.
$\mathbf{Y}_i$	$n_G \times 1$	Collection of Y_{ij} across genes for subject i only.

Table 1: Notation for defining the ICeD-T model.

A.2.2 Overview of Optimization Algorithm of ICeD-T

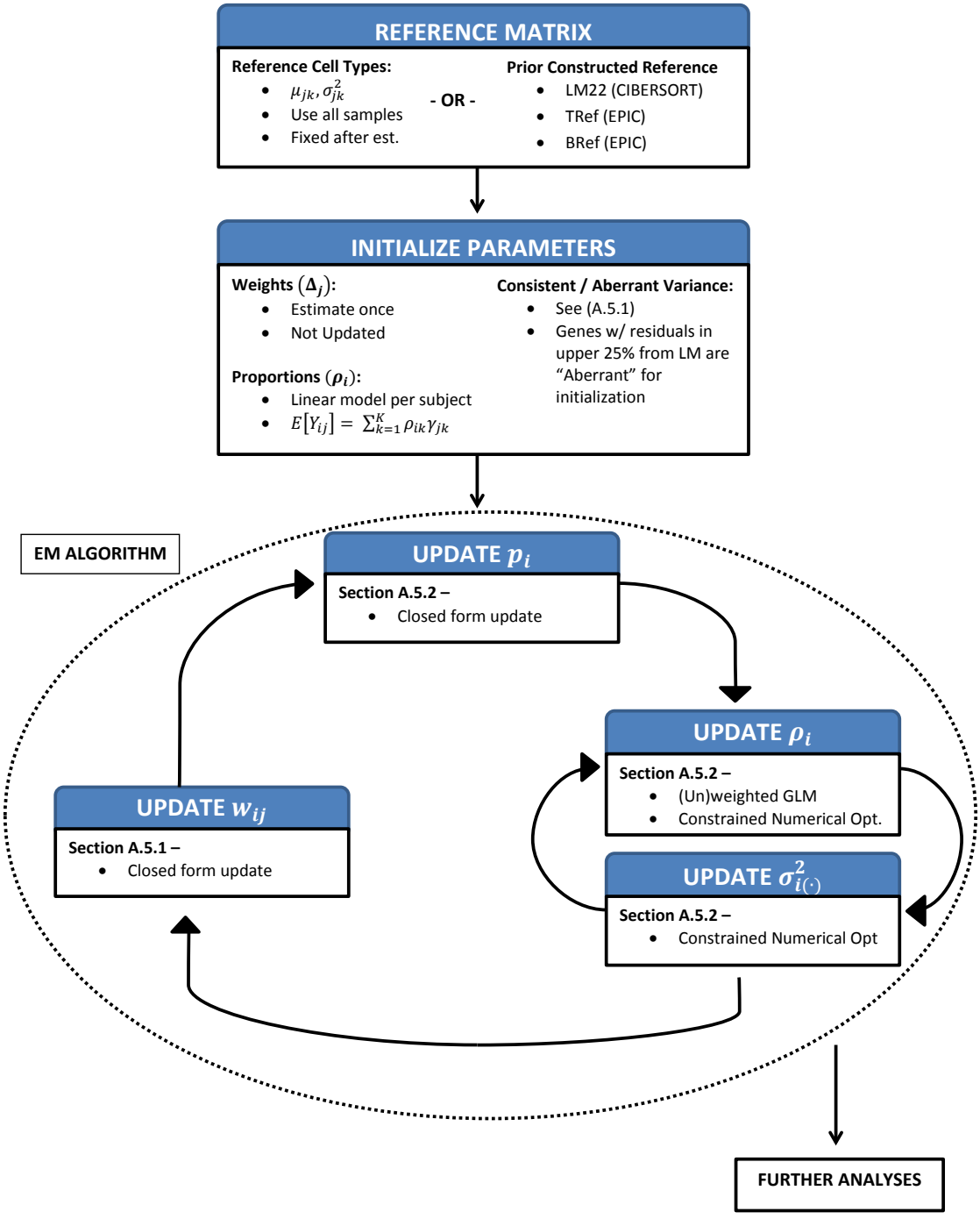


Figure 1: Visual representation of the ICeD-T algorithm from development of reference matrices to EM algorithm.

A.3 Pure Sample Optimization

We focus first on estimation using purified reference samples. Recall that for reference sample h of cell type k, the expression at marker gene j is assumed to follow a log-normal distribution, given by:

$$Z_{jkh} \sim \mathcal{LN}\left(\mu_{jk}, \sigma^2_{jk}\right).$$

The first and second central moments of which are given by:

$$\begin{aligned} E[Z_{jkh}] &= \gamma_{jk} = \exp\left(\mu_{jk} + \sigma^2_{jk}/2\right), \\ V[Z_{jkh}] &= \gamma^2_{jk}\left(\exp\left(\sigma^2_{jk}\right) - 1\right). \end{aligned}$$

Assuming independence of expression across genes within a sample and across samples, the estimators of μ_{jk}, σ^2_{jk} and γ_{jk} are obvious:

$$\begin{aligned} \hat{\mu}_{jk} &= \frac{\sum_{r=1}^{n_q} \log(Z_{jkh})}{n_k}, \\ \hat{\sigma}^2_{jk} &= \frac{\sum_{r=1}^{n_q} \left[\log(Z_{jkh}) - \hat{\mu}_{jk}\right]^2}{n_k - 1}, \\ \hat{\gamma}_{jk} &= \exp\left(\hat{\mu}_{jk} + \hat{\sigma}^2_{jk}/2\right). \end{aligned}$$

In the low sample size setting, we may borrow information across cell types for estimating the variance. We do this in the following way:

$$\hat{\sigma}^2_j = \frac{\sum_{k=1}^{K} \sum_{h=1}^{n_k} \left[\log(Z_{jkh}) - \hat{\mu}_{jk}\right]^2}{n_P - 1},$$

giving

$$\hat{\sigma}^2_{jk} = \left(\frac{n_k}{n_P}\right) \hat{\sigma}^2_{jk} + \left(\frac{n_P - n_k}{n_P}\right) \hat{\sigma}^2_j.$$

A.4 Mixture Sample Optimization

The ICeD-T model assumes that distribution of expression at a single gene in tumor sample i is a mixture over two log-normals, one component assuming the gene is a consistent gene and the other an aberrant one. This distribution is given by:

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

where:

$$\begin{aligned}\tilde{\mu}_{ij(\cdot)} &= \log\left(\sum_{k=1}^{K} \rho_{ik}\gamma_{jk}\right) - \sigma^2_{ij(\cdot)}/2, \\ \sigma^2_{ij(\cdot)} &= \Delta_j \sigma^2_{i(\cdot)}.\end{aligned}$$

ICeD-T can be run using two options. Option (1) represents a homoscedasticity assumption and assumes $\Delta_j = 1$ for all j. Option (2) allows for these variance weights to differ and must be specified before optimization. Utilizing these assumptions for variance provides superior performance in the estimation of cell type proportions compared to a direct application of Fenton-Wilkinson.

We also note that the separating feature between consistent and aberrant genes is the assumed variance. In particular, aberrant genes are assumed to have a larger variance. In essence, a larger variance for aberrant genes "flattens" the observed likelihood, allowing for values inconsistent with the model proportions to become more likely.

In order to optimize this mixture distribution, we utilize an EM algorithm. We introduce missing data in the form of indicators of class membership, H_{ij}. When H_{ij} is 1, the gene is assumed consistent and when H_{ij} is 0, aberrant. Thus, a complete data log-likelihood for subject i is given by:

$$\begin{aligned}\ell_i = \sum_{j=1}^{n_G} H_{ij}\left[\log(p_i) - (1/2)\log(\sigma^2_{ijC}) - \frac{1}{2\sigma^2_{ijC}}\left(\log(y_{ij}) - \tilde{\mu}_{ijC}\right)^2\right] + \\ (1-H_{ij})\left[\log(1-p_i) - (1/2)\log(\sigma^2_{ijA}) - \frac{1}{2\sigma^2_{ijA}}\left(\log(y_{ij}) - \tilde{\mu}_{ijA}\right)^2\right].\end{aligned}$$

The EM algorithm will replace H_{ij} with their posterior expectations $w_{ij} = E\left[H_{ij}\middle|Y_{ij}, \Theta\right]$ prior to optimization at each iteration where Θ is a collection of current estimates of abundances, individual variances, and aberrance proportions.

A.4.1 Update Posterior Means

For a set of parameter estimates, denoted by Θ, it is readily seen that:

$$\begin{aligned} w_{ij} &= E\left[H_{ij}\middle|Y_{ij}=y_{ij},\Theta\right] \\ &= \frac{\left(\frac{p_i}{\sigma_{ijC}}\right)\exp\left\{\left(\frac{-1}{2\sigma^2_{ijC}}\right)\left[\log(y_{ij})-\tilde{\mu}_{ijC}\right]^2\right\}}{\left(\frac{p_i}{\sigma_{ijC}}\right)\exp\left\{\left(\frac{-1}{2\sigma^2_{ijC}}\right)\left[\log(y_{ij})-\tilde{\mu}_{ijC}\right]^2\right\}+\left(\frac{1-p_i}{\sigma_{ijA}}\right)\exp\left\{\left(\frac{-1}{2\sigma^2_{ijA}}\right)\left[\log(y_{ij})-\tilde{\mu}_{ijA}\right]^2\right\}}. \end{aligned}$$

A.4.2 Update p_i, $\sigma^2_{i(\cdot)}$, and $\boldsymbol{\rho}_i$

It is simple to show that the likelihood is separable with respect to p_i and $\left(\boldsymbol{\rho}_i^T,\sigma^2_{iC},\sigma^2_{iA}\right)$. Thus, we may estimate these parameters separately.

Update p_i:
We update p_i with its MLE estimate, given by:

$$\hat{p}_i=\frac{\sum_{j=1}^{n_G}E[H_{ij}|Y_{ij},\Theta]}{n_G}.$$

Update $\left(\boldsymbol{\rho}_i^T,\sigma^2_{iC},\sigma^2_{iA}\right)$
The cell type proportions and variance parameters are not separable within the likelihood and must be updated simultaneously. We opt for a block coordinate ascent algorithm consisting of two blocks; cell type proportions compose block 1 and variance terms compose block 2. Block 1 is updated while block 2 is held fixed, then block 2 is updated while block 1 is fixed. This process is repeated until convergence.

Consider first the variance terms. Holding the cell type proportions fixed, the terms pertaining to aberrant and consistent genes are separable. We focus on the portion of the complete data log-likelihood pertaining to the consistent variance term σ^2_{iC}, though similar results hold for σ^2_{iA}:

$$\begin{aligned} \ell_i\left(\sigma^2_{iC}\right) &= \sum_{j=1}^{n_G} w_{ij}\left[-(1/2)\log\left(\Delta_j\sigma^2_{iC}\right)-\frac{1}{2\Delta_j\sigma^2_{iC}}\left(\log(y_{ij})-\tilde{\mu}_{ijC}\right)^2\right] \\ &= \sum_{j=1}^{n_G} w_{ij}\left[-(1/2)\log\left(\Delta_j\sigma^2_{iC}\right)-\frac{1}{2\Delta_j\sigma^2_{iC}}\left(\nu_{ij}+\Delta_j\sigma^2_{iC}/2\right)^2\right] \end{aligned}$$

where $\nu_{ij}=\log\left(y_{ij}\right)-\log\left(\sum_{k=1}^{K}\rho_{ik}\gamma_{jk}\right)$.

Taking the first derivative with respect to the consistent variance term, we have:

$$\dot{\ell}\left(\sigma^2_{iC}\right)=\sum_{j=1}^{n_G} w_{ij}\left[\left(\frac{-1}{2\sigma^2_{iC}}\right)+\left(\frac{1}{2\Delta_j\sigma^4_{iC}}\right)\left(\nu_{ij}+\Delta_j\sigma^2_{iC}/2\right)^2-\left(\frac{1}{2\sigma^2_{iC}}\right)\left(\nu_{ij}+\Delta_j\sigma^2_{iC}/2\right)\right].$$

Under option (2) where Δ_j varies across j's, we did not find a closed form update for these variance terms and thus opted to use numerical optimization. Under option (1) where $\Delta_j = 1$ for all j's, we can further reduce this equation by plugging in $\Delta_j = 1$:

$$\begin{aligned}
\dot{\ell}\left(\sigma_{iC}^2\right) &= \sum_{j=1}^{n_G}\left(\frac{w_{ij}}{2\sigma_{iC}^4}\right)\left[-\sigma_{iC}^2 + \nu_{ij}^2 + \nu_{ij}\sigma_{iC}^2 + \sigma_{iC}^4/4 - \sigma_{iC}^2\nu_{ij} - \sigma_{iC}^4/2\right] \\
&= \sum_{j=1}^{n_G}\left(\frac{w_{ij}}{2\sigma_{iC}^4}\right)\left[-\left(\sigma_{iC}^4/4 + \sigma_{iC}^2\right) + \nu_{ij}^2\right] \\
&= \sum_{j=1}^{n_G}\left(\frac{w_{ij}}{2\sigma_{iC}^4}\right)\left[-\left(\sigma_{iC}^2/2 + 1\right)^2 + \nu_{ij}^2 + 1\right].
\end{aligned}$$

Setting $\dot{\ell}\left(\sigma_{iC}^2\right)$ to be 0 and solving for σ_{iC}^2, we have a closed form update for σ_{iC}^2:

$$\sigma_{iC}^2 = 2\left[\sqrt{\frac{\sum_{j=1}^{n_G} w_{ij}\nu_{ij}^2}{\sum_{j=1}^{n_G} w_{ij}} + 1} - 1\right].$$

We now turn to the cell type proportion piece, assuming the variance terms are held fixed. The complete data log-likelihood pertaining to these parameters is given by:

$$\begin{aligned}
\ell_i = \sum_{j=1}^{n_G} w_{ij}\left[\log(p_i) - (1/2)\log(\sigma_{iC}^2) - \frac{1}{2\sigma_{iC}^2}\left(\log(y_{ij}) - \tilde{\mu}_{ijC}\right)^2\right] + \\
(1 - w_{ij})\left[\log(1 - p_i) - (1/2)\log(\sigma_{iA}^2) - \frac{1}{2\sigma_{iA}^2}\left(\log(y_{ij}) - \tilde{\mu}_{ijA}\right)^2\right].
\end{aligned}$$

Before constructing the derivative of this likelihood with respect to our cell type proportions, we examine the derivatives of an interior term of the likelihood to improve clarity of the full derivation. In the following, let $\eta_{ij} = \sum_{k=1}^{K}\rho_{ik}\gamma_{jk}$.

$$\frac{\partial \mu_{ij(\cdot)}}{\partial \rho_i} = \frac{\partial}{\partial \rho_i}\left[\log\left(\sum_{k=1}^{K}\rho_{ik}\gamma_{jk}\right) - \sigma_{ij(\cdot)}^2/2\right] = \frac{\boldsymbol{\gamma}_j}{\sum_{k=1}^{K}\rho_{ik}\gamma_{jk}} = \frac{\boldsymbol{\gamma}_j}{\eta_{ij}}$$

Plugging this value into the gradient for the complete data log-likelihood, we have:

$$\begin{aligned}
\dot{\ell}_i &= \sum_{j=1}^{n_G} w_{ij}\left[(1/\sigma_{ijC}^2)\left(\log(y_{ij}) - \tilde{\mu}_{ijC}\right)\right]\left(\frac{\partial \mu_{ij}^{(C)}}{\partial \rho_i}\right) + \\
&\qquad (1 - w_{ij})\left[(1/\sigma_{ijA}^2)\left(\log(y_{ij}) - \tilde{\mu}_{ijA}\right)\right]\left(\frac{\partial \mu_{ij}^{(A)}}{\partial \rho_i}\right) \\
&= \sum_{j=1}^{n_G}\boldsymbol{\gamma}_j\left[\left(\frac{w_{ij}\left(\log(y_{ij}) - \tilde{\mu}_{ijC}\right)}{\sigma_{ijC}^2\eta_{ij}}\right) + \left(\frac{(1 - w_{ij})\left(\log(y_{ij}) - \tilde{\mu}_{ijA}\right)}{\sigma_{ijA}^2\eta_{ij}}\right)\right].
\end{aligned}$$

To ensure proper constraints during fit, numerical optimization routines from R's `auglag` function in the `alabama` package are used to optimize the log-likelihood with respect to ρ_i.

When no information is assumed for the proportion of a $(K+1)$-st cell type (e.g. a tumor cell type), these proportions are non-negative and allowed to sum to a value less than 1. If the proportion of a $(K+1)$-st cell type is assumed (e.g. tumor purity), the proportions are constrained to sum to $1-\rho_{K+1}$. As noted in the main paper, the $(K+1)$-st cell type is assumed not to express or to express at a minimal level across the n_G genes used for optimization.

B Simulations Supplement

The first assessment of the estimation properties of the ICeD-T model was performed on *in silico* simulated datasets. For each simulation, we constructed two sets of expression pseudo-experiments: reference expression datasets from 5 simulated reference cell types and reference expression datasets from 135 mixture datasets composed of expression from 4 of these 5 cell types. Each expression experiment consists of expression values across 250 common loci (genes). Within the mixtures, one cell type represents a "missing" cell type for each sample; this cell type is known to be present in the mixture but it does not express at the 250 modeled loci.

These simulations were built in three main steps: Step (1) generates purified reference sample expressions and variance measures; Step (2) generates mixture expression files for deconvolution; and step (3) edits the output expressions from step 2 to allow for aberrant gene behavior.

B.1 Step 1 - Generating Pure Sample Expressions

The first element in generating pure sample expressions is to define profiles from which each "purified reference" sample is simulated. For each locus separately, it is randomly determined whether the locus is lowly, moderately, or highly expressed. In addition, one of the four expressed cell types is labeled the indicated cell type for this locus while the remaining cell types are considered background. We then simulate a mean log-expression for each gene and cell type according to the following table:

Level	Pct. Loci	Background	Indicated
Low	33%	$Uniform(2.0, 4.0)$	$Uniform(3.5, 5)$
Moderate	33%	$Uniform(4.0, 6.0)$	$Uniform(5.5, 7)$
High	33%	$Uniform(6.0, 8.0)$	$Uniform(7.5, 9)$

Once the mean log-expressions are simulated, we must construct a reasonable variance schema for these average log-expression profiles. We construct a mean-variance relationship in the log-expression setting by mirroring an example found in FPKM-normalized RNA-seq data.

Read counts from purified samples of B-cells (20), CD4 T-cells (20), CD8 T-cells (20), Monocytes (20), Neutrophils (20) and Natural Killers (14) were downloaded from the Array Express website from the Linsley et al study [3]. The read counts for each sample are FPKM normalized, utilizing the (75th-percentile read count/1000) instead of total read depth for each subject. The mean-variance relationship is modeled across 441 immune-related genes for each of these six cell types using a Loess curve, similar to the procedure utilized by VOOM [4]. This Loess curve was used to map the simulated log-expression means for each gene and cell type to a data-supported variance measure. Random error was also introduced.

Following the generation of the mean and variance profiles, the 5 or 15 purified, reference-sample pseudo-experiments are generated for each cell type from its profile via a log-normal distribution.

B.2 Step 2 - Generating Mixture Expressions

We must now generate the mixture expression pseudo-experiments. We first generate the proportion of the missing cell type from a standard normal distribution with mean 0.60 and standard deviation 0.15. In addition, any of these proportions falling below 17% or above 95% are set at

17% and 95% respectively. The remaining proportions are simulated from a Dirichlet distribution with average abundances ranging from 15% to 40%.

With the proportions generated for each of the 5 cell types and each subject, we turn to simulating the expression experiments. For each subject individually, we construct mixture experiments according to the following algorithm.

(1) Simulate cell type-specific expression for non-missing cell types

$$X_{ijk} \sim \exp\left(\mathcal{N}\left(\mu_{jk}, \sigma^2_{jk}\right)\right)$$

(2) Mix cell type-specific expression

$$Y_{ij} \sim \sum_{k=1}^{4} \rho_{ik} X_{ijk}$$

Thus, these mixture expression experiments are simulated as true convolutions of independent log-normals. In this way, we can examine the adequacy of our approximated distribution.

B.3 Step 3 - Edit Mixtures to Create Aberrance

The final step in the mixture experiments is to allow loci to misbehave. We allow 3 mechanisms for misbehavior. Mechanism 1 takes the expression of the indicated cell type and downregulates it to 25% - 75% of its true level; mechanism 2 takes the expression of the indicated cell type and upregulates it to 133% - 400% of its true level; and mechanism 3 allows the missing cell type to express at the background levels established above. The table below summarizes these mechanisms.

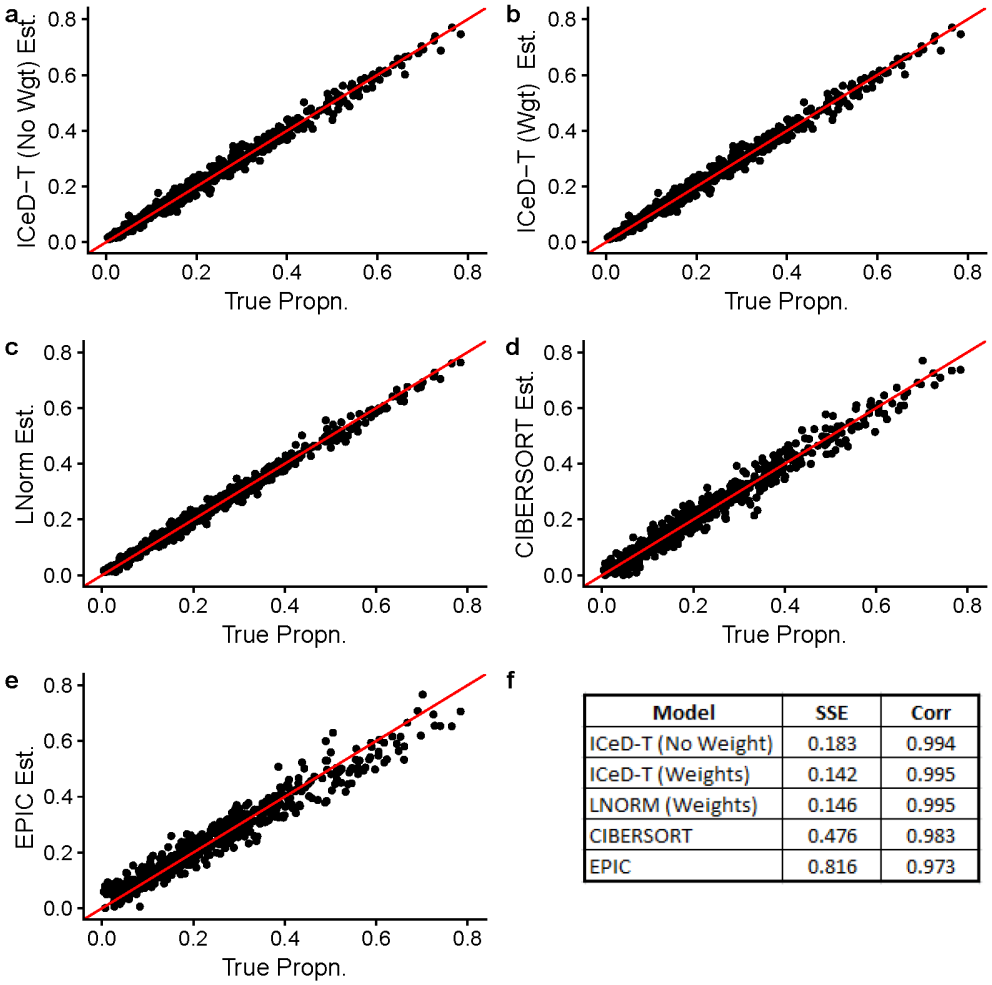
Mechanism	**Pct. Ab. Loci**	**Indicated CT Effect**	**Missing CT Exp.**
1 - Downregulate	25%	$Uniform(25\%, 75\%)$	0
2 - Upregulate	25%	$Uniform(133\%, 400\%)$	0
3 - Missing Exp.	50%	0	$Uniform(.,.)$

Table 2: Pct Ab. Loci = Percentage of Aberrant Loci Effected, Indicated CT Effect = Effect on the expression of Indicated Cell Type, Missing CT Exp = Expression Level of Missing Cell Type

For impacted loci, expression is resimulated as in B.2 with the revised expression profiles. The number of loci impacted is allowed to vary from 0% to 30% of the total expression and the resulting estimates are examined.

B.4 Additional Simulation Results

Pct Ab. = 0%, No. Rep. = 5

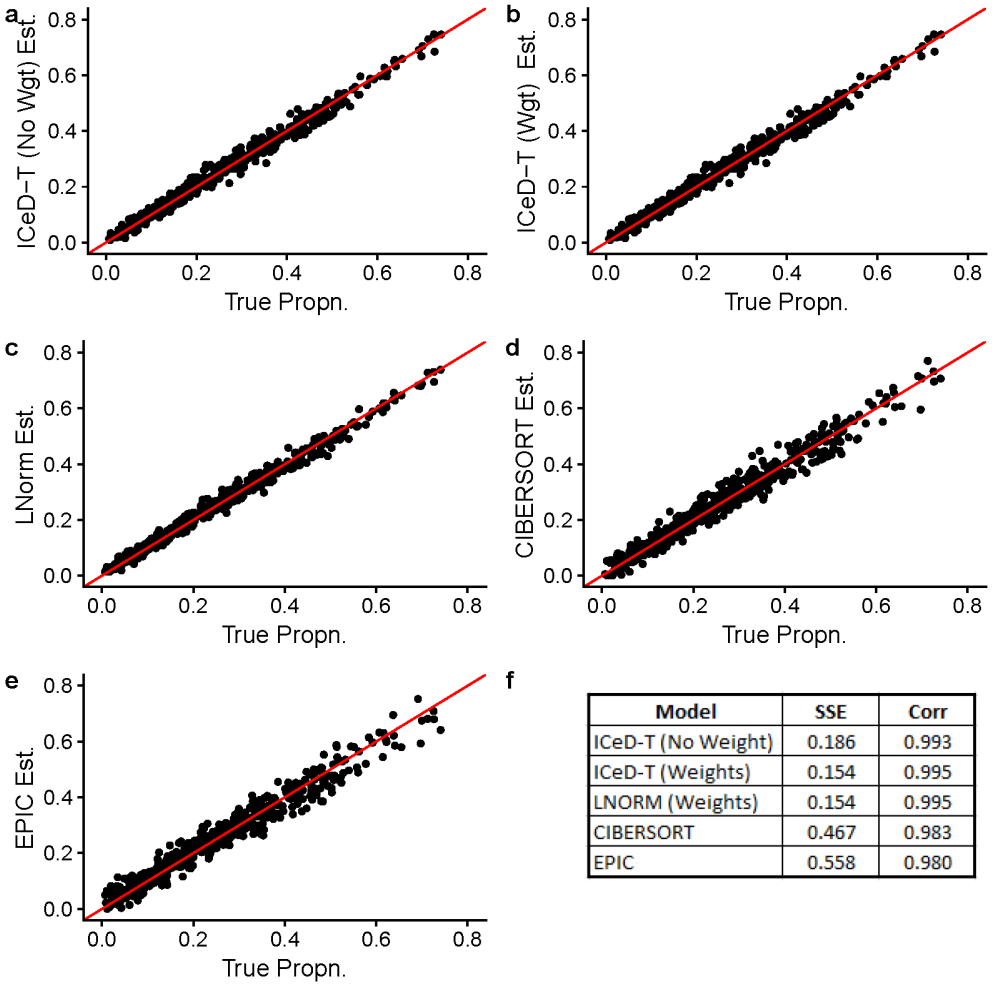


f

Model	SSE	Corr
ICeD-T (No Weight)	0.183	0.994
ICeD-T (Weights)	0.142	0.995
LNORM (Weights)	0.146	0.995
CIBERSORT	0.476	0.983
EPIC	0.816	0.973

Figure 2: Visualizing simulation results with 5 reference samples per cell type and no aberrance.

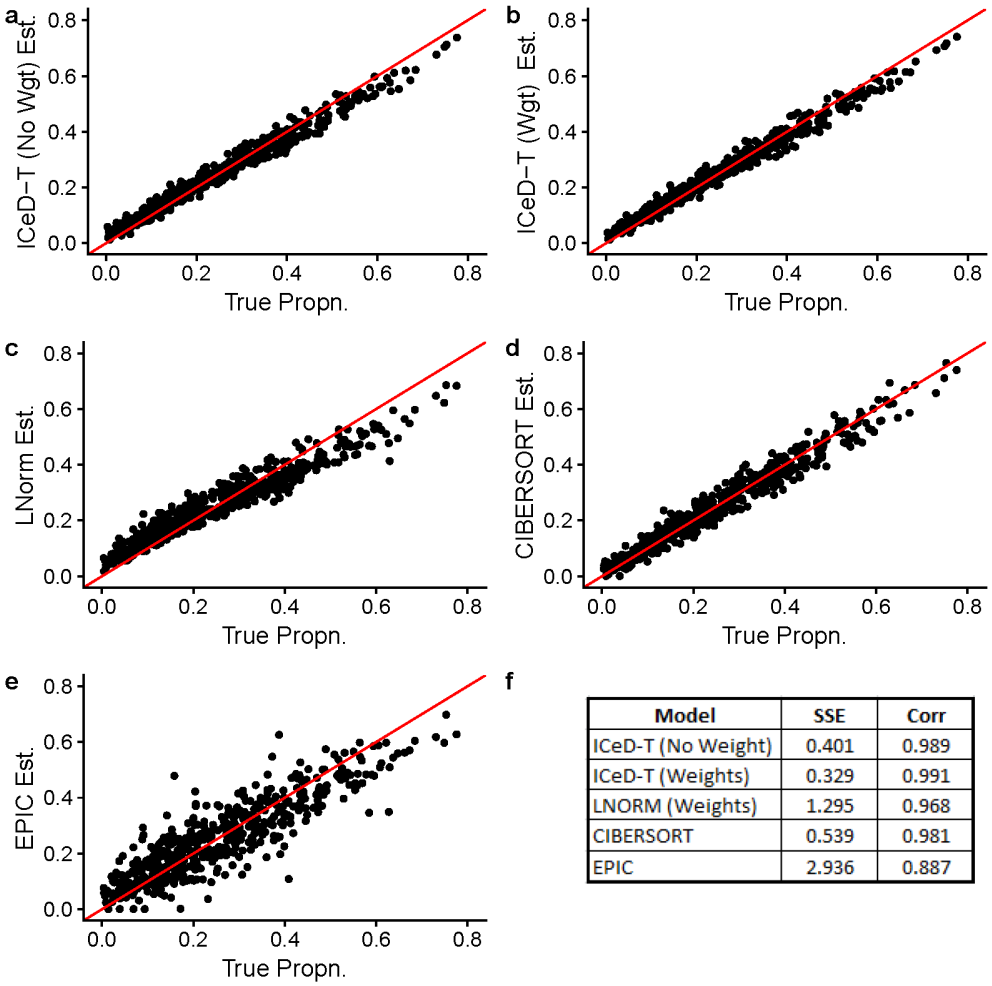
Pct Ab. = 0%, No. Rep. = 15



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

Figure 3: Visualizing simulation results with 15 reference samples per cell type and no aberrance.

Pct Ab. = 15%, No. Rep. = 5



f

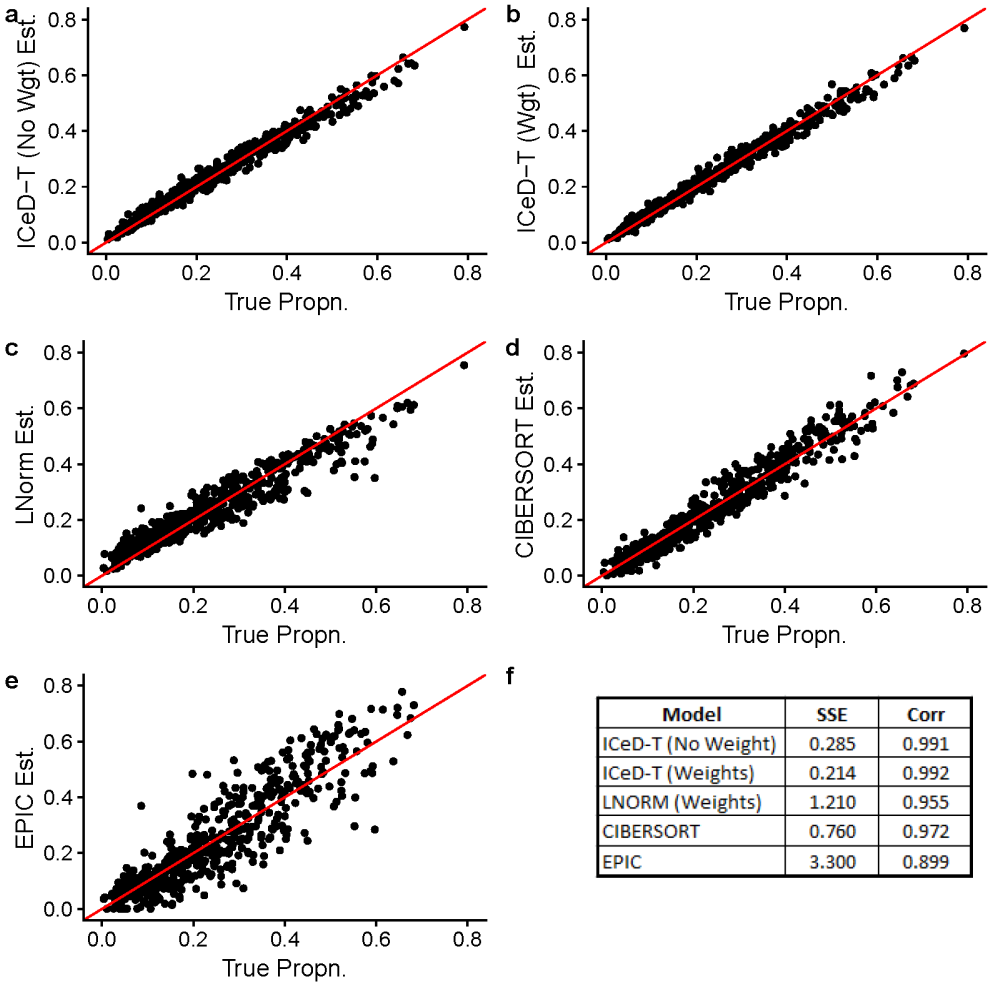
Model	SSE	Corr
ICeD-T (No Weight)	0.401	0.989
ICeD-T (Weights)	0.329	0.991
LNORM (Weights)	1.295	0.968
CIBERSORT	0.539	0.981
EPIC	2.936	0.887

Figure 4: Visualizing simulation results with 5 reference samples per cell type and 15% of genes behaving aberrantly.

Model	**Aberrant**	**1Q**	**Med**	**3Q**
ICeD-T (No Weight)	Yes	0.000	0.114	0.607
	No	0.625	0.748	0.824
	p_i	0.572	0.612	0.655
ICeD-T (Weights)	Yes	0.004	0.468	0.823
	No	0.804	0.884	0.931
	p_i	0.718	0.768	0.803

Table 3: Summarizing ICeD-T's ability to detect aberrant gene behavior (Pct. Ab. = 15%, No. Rep. = 5).

Pct Ab. = 18%, No. Rep. = 15



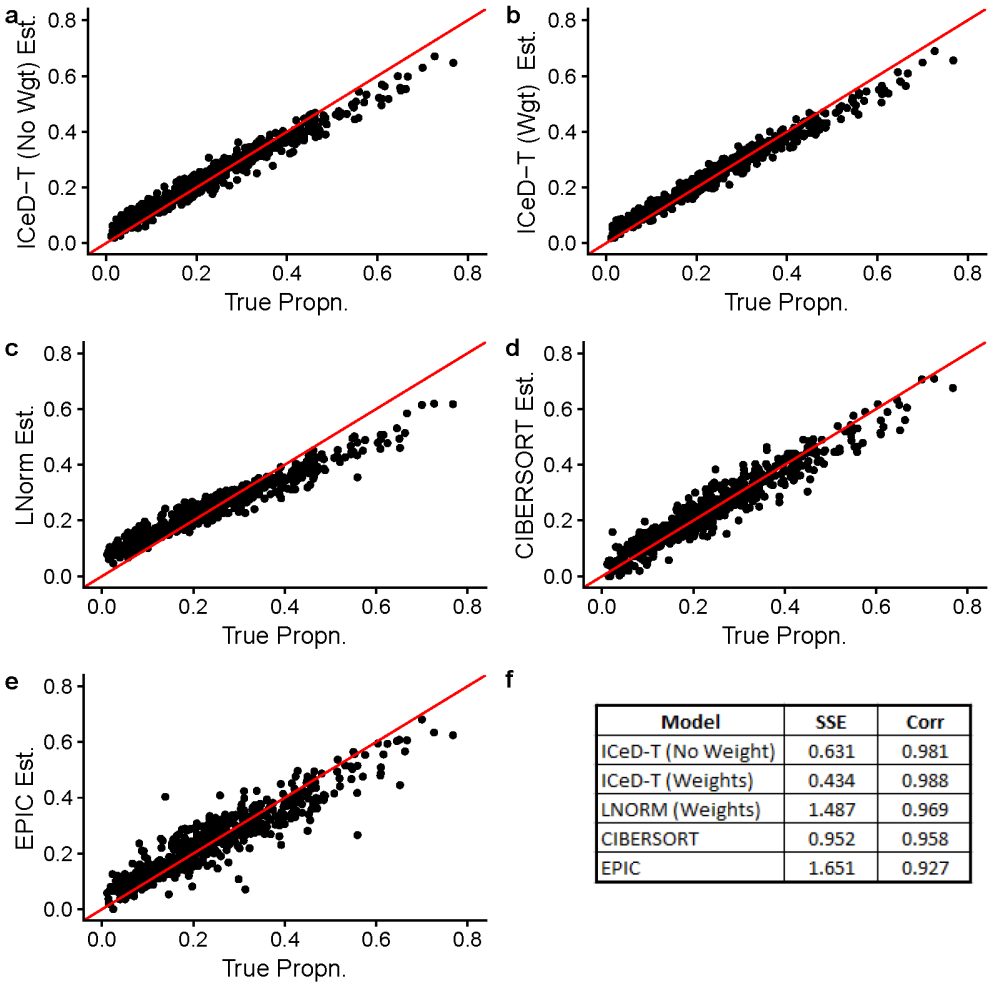
Model	SSE	Corr
ICeD-T (No Weight)	0.285	0.991
ICeD-T (Weights)	0.214	0.992
LNORM (Weights)	1.210	0.955
CIBERSORT	0.760	0.972
EPIC	3.300	0.899

Figure 5: Visualizing simulation results with 15 reference samples per cell type and 18% of genes behaving aberrantly.

Model	**Aberrant**	**1Q**	**Med**	**3Q**
ICeD-T (No Weight)	Yes	0.000	0.043	0.538
	No	0.647	0.769	0.838
	p_i	0.579	0.613	0.657
ICeD-T (Weights)	Yes	0.000	0.114	0.744
	No	0.797	0.872	0.920
	p_i	0.697	0.738	0.772

Table 4: Summarizing ICeD-T's ability to detect aberrant gene behavior (Pct. Ab. = 18%, No. Rep. = 15).

Pct Ab. = 30%, No. Rep. = 5



f

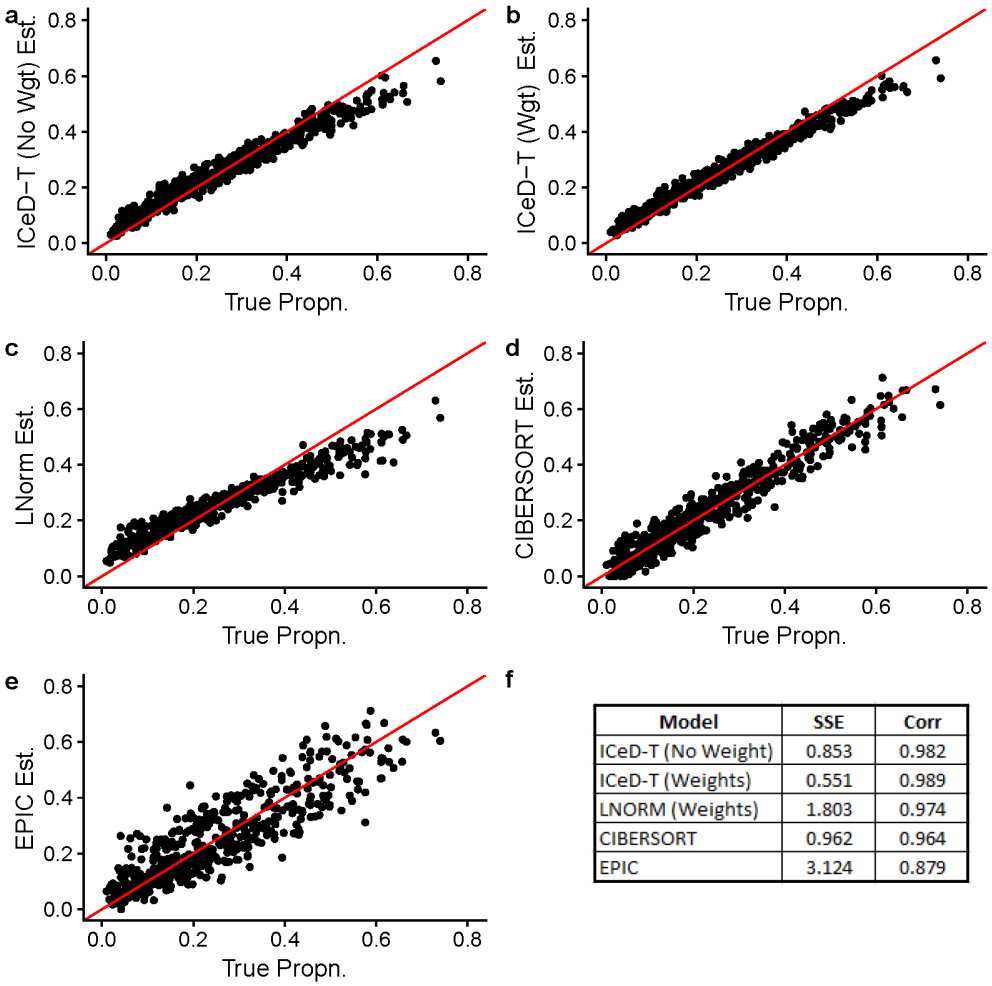
Model	SSE	Corr
ICeD-T (No Weight)	0.631	0.981
ICeD-T (Weights)	0.434	0.988
LNORM (Weights)	1.487	0.969
CIBERSORT	0.952	0.958
EPIC	1.651	0.927

Figure 6: Visualizing simulation results with 5 reference samples per cell type and 30% of genes behaving aberrantly.

Model	**Aberrant**	**1Q**	**Med**	**3Q**
ICeD-T (No Weight)	Yes	0.001	0.194	0.645
	No	0.618	0.659	0.791
	p_i	0.530	0.555	0.587
ICeD-T (Weights)	Yes	0.011	0.552	0.824
	No	0.780	0.862	0.897
	p_i	0.673	0.707	0.725

Table 5: Summarizing ICeD-T's ability to detect aberrant gene behavior (Pct. Ab. = 30%, No. Rep. = 5).

Pct Ab. = 35%, No. Rep. = 15



Model	SSE	Corr
ICeD-T (No Weight)	0.853	0.982
ICeD-T (Weights)	0.551	0.989
LNORM (Weights)	1.803	0.974
CIBERSORT	0.962	0.964
EPIC	3.124	0.879

Figure 7: Visualizing simulation results with 15 reference samples per cell type and 35% of genes behaving aberrantly.

Model	Aberrant	1Q	Med	3Q
ICeD-T (No Weight)	Yes	0.002	0.202	0.582
	No	0.574	0.678	0.734
	p_i	0.480	0.503	0.529
ICeD-T (Weights)	Yes	0.013	0.406	0.730
	No	0.714	0.790	0.837
	p_i	0.597	0.621	0.643

Table 6: Summarizing ICeD-T's ability to detect aberrant gene behavior (Pct. Ab. = 30%, No. Rep. = 15).

We see from the above that the ICeD-T model with and without weights provides the best fit for these simulated data in terms of both sum of squared error and correlation. The aberrance model adequately handles the misbehavior across loci even up to 30% aberrance, with the weighted model providing the strongest estimation. It most closely estimates the proportion of aberrant genes and provides stronger distinctions in the probabilities of aberrance given the data.

As we reach 30% aberrance, we do note a slight bias in ICeD-T's results beginning to become evident near the tails. However, even when compared against CIBERSORT–a method which provides a very strong runner-up in these simulated data– we see that ICeD-T is superior. This is a classic case of the bias-variance trade-off. ICeD-T allows some bias to impact results as aberrance increases, but maintains a strong linear relationship. CIBERSORT, on the other hand, experiences increasing variability and a slightly diminished linear relationship as the amount of aberrance increases.

We also fit the ICeD-T model without using estimates of tumor purity (data not shown). The model performs well up to 30% aberrance, however, at around 30% aberrance it begins to struggle to capture aberrant genes appropriately. Regardless of this fact, the ICeD-T model with weights continues to be one of the strongest performers even up to 30% aberrance.

B.5 Robustness to different initial values of cell type compositions

It is important to evaluate whether our algorithm is robust to different initial values of cell type compositions. Robustness to initial values would also suggest that the likelihood surface is concave or very close to be concave, and thus there is little risk to reach wrong estimates of local maximizer. We randomly picked one simulated bulk tumor sample with tumor purity around 45%, and assumes there are five cell types including tumor cells. Then we generated 1,000 initial values of cell type compositions from Dirichlet distribution with parameters $\alpha = (1, 1, 1, 1, 1)^T$, hence without any prior information on cell type composition. As shown in the upper panels of Figure 8, initial values of cell type proportions are indeed randomly simulated. However, despite such large divergence of initial values, the final estimates of cell type proportions all converge to similar values across the 1,000 replicates. When given tumor purity, there are virtually no difference in the final cell type composition estimates. Without tumor purity information, the final estimates have very slight difference.

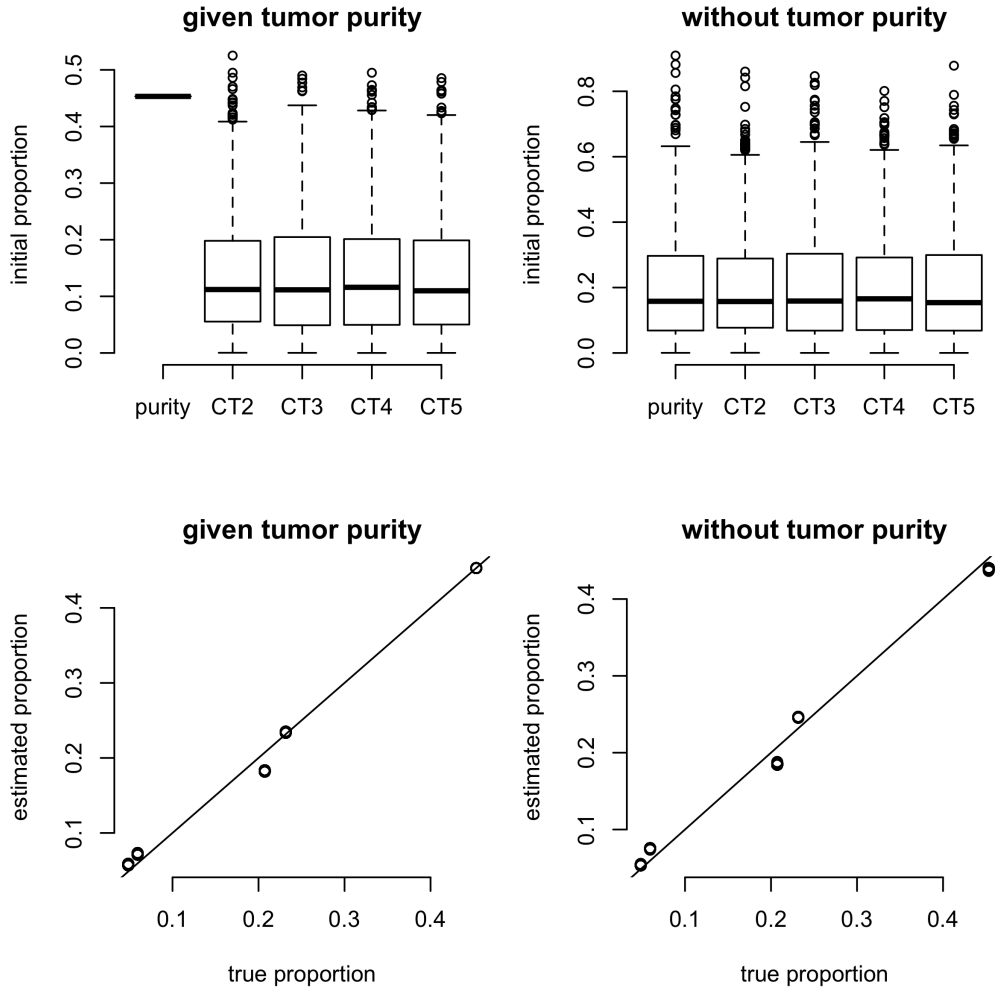


Figure 8: Evaluation of the robustness of the ICeD-T algorithm with 1,000 different initial values of cell type proportions. Upper panels: initial values of cell type proportions, which are uniformly distributed. Lower panels: final estimates of cell type compositions.

C CIBERSORT Flow Cytometry Validation

The second assessment of the performance properties of the ICeD-T model is performed in real data. In their paper "Robust Enumeration of Cell Subsets from Tissue Expression Profiles," the creators of CIBERSORT validate their modeling procedure on peripheral blood mononuclear cells (PBMCs) extracted from 20 adult subjects. We reanalyze this dataset using CIBERSORT's web application, the ICeD-T model, and EPIC.

C.1 Data

PBMCs were extracted from each of 20 adult subjects. For each sample, expression profiles were created using microarray expression analysis. Additionally, each sample was examined using flow cytometry to measure the ground-truth abundance of each of the immune cell types composing the PBMCs. The resulting datasets were provided to us directly by Newman et al. In addition, the microarray expression data from purified samples of 22 immune cell types used to construct LM22 were also provided.

C.2 Cell Type Size Correction

The authors of EPIC advocate the use of cell size factors to correct regression results for differences in the productivity of various cell types composing mixture experiments. In their work, "Simultaneous Enumeration Of Cancer And Immune Cell Types From Bulk Tumor Gene Expression Data," they note that cells of various types produce differing levels of mRNA. We borrow these cell size factors here and use them to correct the results of CIBERSORT and ICeD-T as was performed below. The cell size factors utilized here are provided below. Cell size factors are incorporated into

Cell	Size Factor	Extensions
B-Cells	0.40	Naive and memory B-cells
T-Cells	0.40	Naive, memory-resting and memory-activated CD4 T-cells; CD8 T-cells; Delta-Gamma T-cells
NK cells	0.42	None
Monocytes	1.40	Macrophages, Dendritic Cells
Neutrophils	0.15	Eosinophils, Mast Cells

Table 7: EPIC-derived cell type size factors with extensions to cell types not explicitly measured.

model estimates after running the ICeD-T or CIBERSORT models as was done in EPIC. Define s_k to be the cell size factor for a cell type k. Then the revised estimate of abundance for cell type k is given by:

$$\rho_k^* = \frac{\rho_k / s_k}{\sum_{i=1}^{K} (\rho_k / s_k)}.$$

C.3 Model Fit Description

CIBERSORT:
The CIBERSORT web application (Version: CIBERSORT Jar 1.06) was used to fit these microarray data. The model was fit using the LM22 signature matrix run with quantile normalization and 500 permutations.

EPIC:
The EPIC library was downloaded from https://github.com/GfellerLab/EPIC in February 2018. The mixture expression data is quantile normalized and fit to the LM22 reference matrix using EPIC with default options, except scaleExprs set to FALSE.

ICeD-T:
The ICeD-T model was fit to the LM22 reference without specifying the proportions of extraneous cell types in the model and no weights, maximal variance weights, and maximal expression variance weights. Variance weights were computed using the variance of log-transformed expression across all purified references of a given cell type.

Quantile Normalization:
EPIC and ICeD-T require that the modeled mixture data be measured on the same scale as the design matrix utilized for modeling. To this end, the purified references used to compose the LM22 matrix are quantile normalized. The mixture data are then quantile normalized to the target distribution specified by the purified references using the `preprocessCore` library and its functions `normalize.quantiles.determine.target` and `normalize.quantiles.use.target`. This normalization is performed prior to specification of gene and cell type variance measures.

Result Renormalization:
Results are handled in the manner suggested by Newman et al in personal correspondence as was performed for their manuscript. All estimated cell type proportions are restricted to the ten examined cell types: B-cells naive, B-cells memory, CD8+ T-cells, naive CD4+ T-cells, resting memory CD4+ T-cells, activated memory CD4+ T-cells, Delta-gamma T-cells, Activated and resting natural killer cells, and Monocytes (including the modeled macrophage populations). These proportions are then renormalized to sum to 100.

C.4 Fit Comparison

The following table details the correlations and sum of squared errors for each of the fit models. As noted above, each of these measures use cell size corrected proportions for examination.

Model	SSE	Cor
ICeD-T (No Wgt)	13099.93	0.53
ICeD-T (Max Var Wgt)	12050.67	0.59
CIBERSORT	14146.59	0.65
EPIC	29427.74	0.31

Table 8: Fit summary statistics for each model compared against flow cytometry measured ground-truth.

We note from the above that the CIBERSORT model provides the best results in terms of correlations. However, each of the fit ICeD-T models provide superior fit in terms of sums of squared errors. When using variance weights, the correlations between ICeD-T estimates and CIBERSORT become comparable as well ($\sim$0.60 vs. 0.65). Thus, it appears that the ICeD-T method is comparable to CIBERSORT in terms of correlation and provides superior results in terms of error.

In the following considerations, we will focus on the ICeD-T model with maximal variance weights. Despite the fact that the maximal expression weights produced the best fit for both overall correlation and sum of squared errors, it has notably weaker fit for many important cell types (e.g. CD4, CD8). Compared to CIBERSORT, in addition to having lower overall error, ICeD-T appears to provide superior performance for memory B-cells, naive CD4 T-cells, and gamma-delta T-cells. Both CIBERSORT and ICeD-T provide comparable performance with respect to monocyte expression. Both models struggle with CD8 T expression despite being well correlated for this cell type as CIBERSORT tends to overestimate expression by in the upper tail where ICeD-T seems to underestimate.

The results provided by the EPIC model are very poor for this dataset. However, this is not a condemnation of EPIC's use in real data. EPIC was designed for RNA-seq data, not for microarrays. Thus, the weighting structure and gene selection for the fit shown here may not be suitable for EPIC's off-the-shelf options.

ICeD-T, No Weights

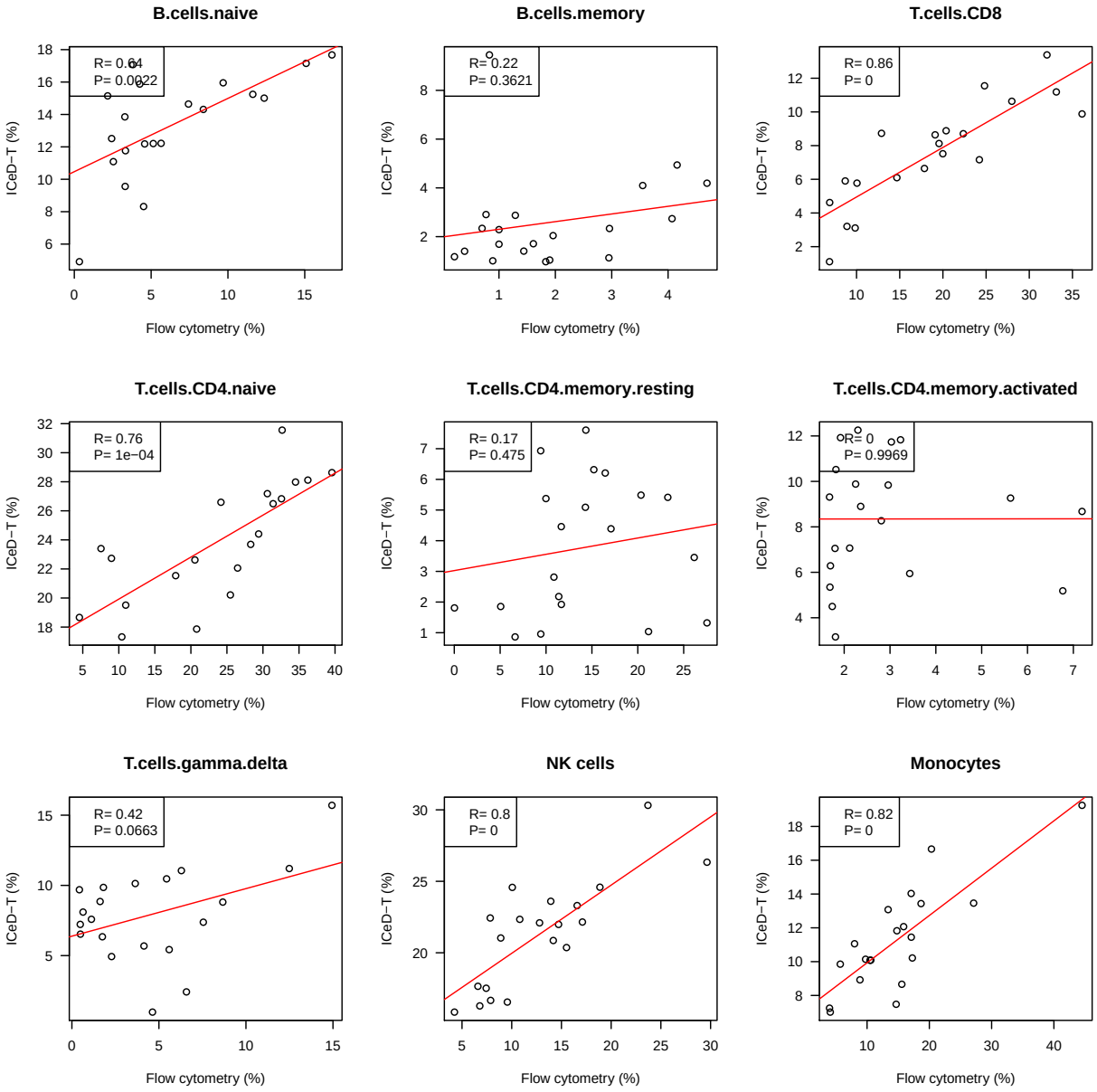


Figure 9: Plotting true, relative abundances of 9 immune cell subpopulations against ICeD-T (no weights) estimates.

ICeD-T, Max Variance Weights

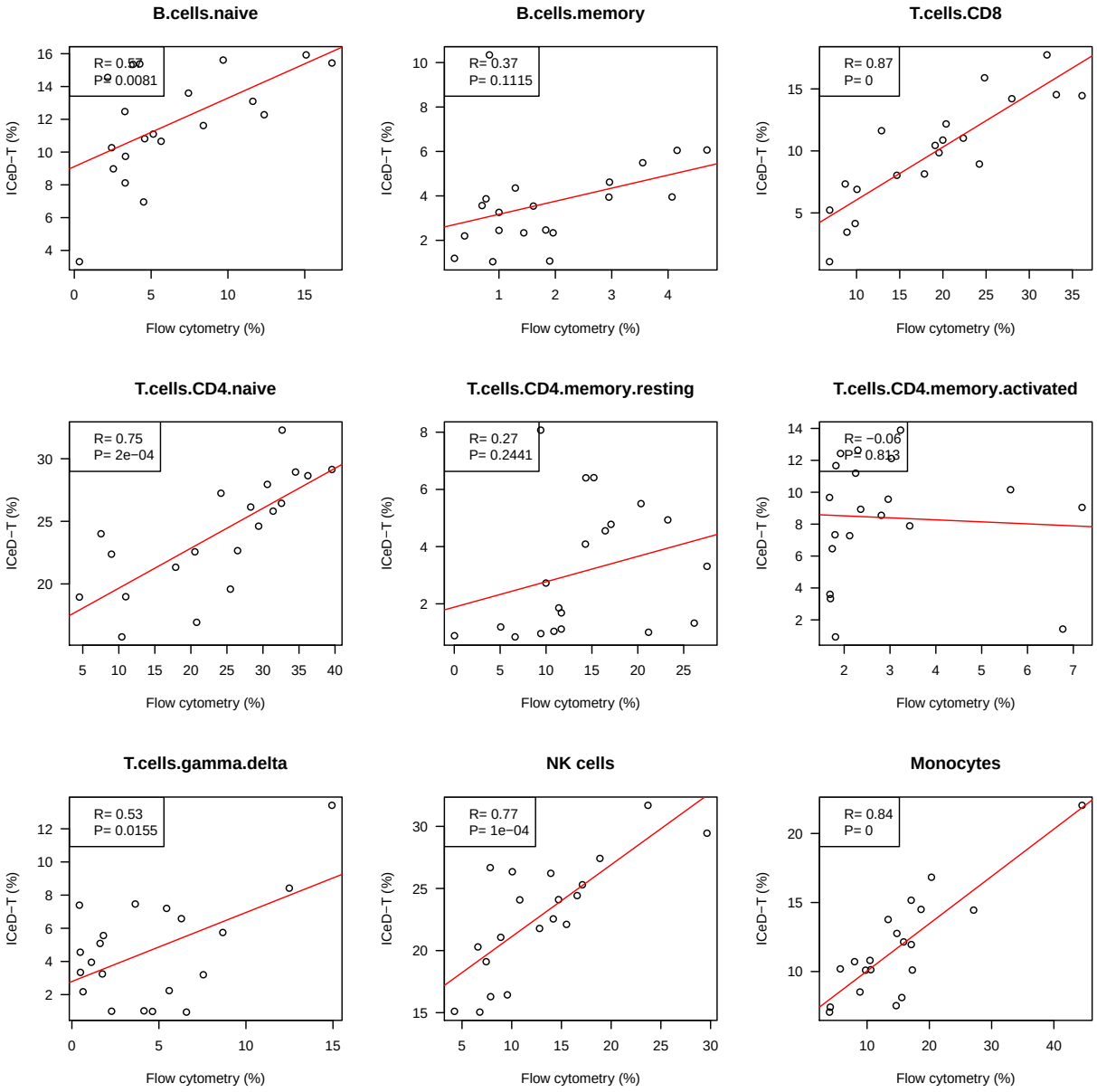


Figure 10: Plotting true, relative abundances of 9 immune cell subpopulations against ICeD-T (weights) estimates.

CIBERSORT

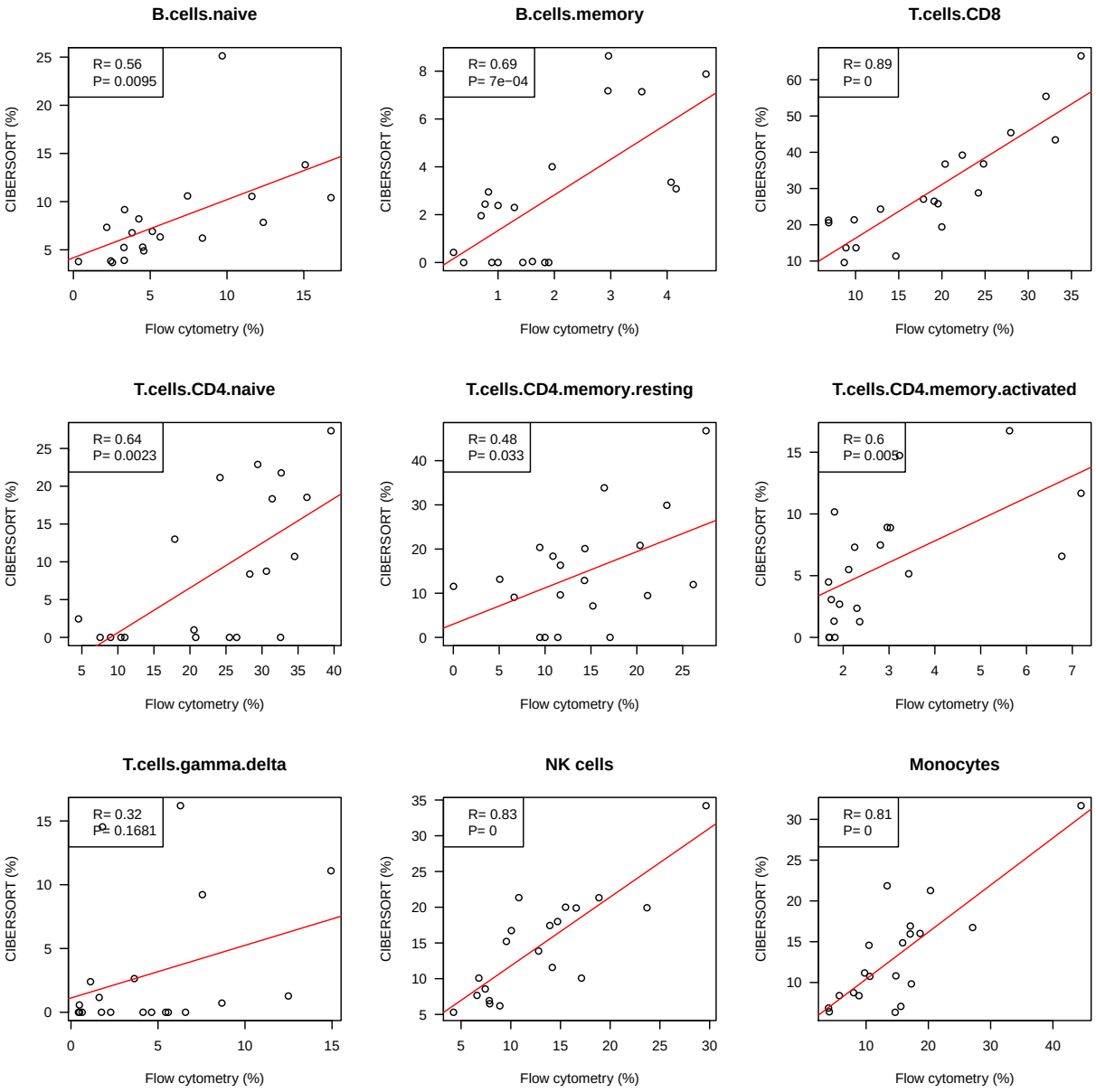


Figure 11: Plotting true, relative abundances of 9 immune cell subpopulations against CIBERSORT estimates.

D EPIC Melanoma Data Validation

The third examination of the estimation properties of ICeD-T is performed on validation data provided by Racle et al. It offers an opportunity to evaluate the performance of ICeD-T on RNA-seq experiments from tumor samples.

D.1 Data

For more information regarding this dataset, see "Simultaneous Enumeration of Cancer and Immune Cell Types from Bulk Tumor Gene Expression Data" from Racle et al. In brief, cells were extracted from the lymph nodes of four patients with stage III melanomas. A portion of each of the single cell suspensions obtained from these subjects was used for a flow cytometric analysis while the remaining portion was used for bulk RNA-sequencing.

The data was extracted from the EPIC library, with path `EPIC-master/data/melanoma_data.rda`. This RData files contains a single list object `melanoma_data`, which houses fields containing the TPM-normalized RNA-seq expression for each subject, the flow-cytometry measured cell type proportions, and the predicted EPIC cell type proportions obtained using the TRef reference matrix.

D.2 Model Fit Description

CIBERSORT:
The CIBERSORT web application (Version: CIBERSORT Jar 1.06) was used to fit these TPM normalized RNA-seq data. The model was fit using the LM22 signature matrix and run with quantile normalization disabled.

EPIC:
The EPIC model was fit to these TPM normalized RNA-seq data using its TRef reference matrix and all default options.

ICeD-T:
The ICeD-T model is fit using all 4 combinations of the following options: (1) Use Tumor Purity: Yes or no? (2) Use maximal variance weights: Yes or No?. For the purposes of this analysis, tumor purity is obtained from the flow cytometry results by combining the proportions of cancer cells and other cells.

D.3 Fit Results

For the results shown below, all immune content is corrected for cell size and renormalized so that proportions are computed with respect to the immune cells in the mixture (B-cells, CD4+ T-cells, CD8+ T-cells, and Natural Killers).

TRUTH	B-cells	CD4+ T	CD8+ T	NK
LAU125	0.9156	0.0414	0.0177	0.0253
LAU1255	0.4639	0.2212	0.3013	0.0136
LAU1314	0.6704	0.2607	0.0652	0.0036
LAU335	0.5271	0.3757	0.0944	0.0028

Table 9: Melanoma Data - True relative proportions of Immune cells

EPIC	B-cells	CD4+ T	CD8+ T	NK	SSQ
LAU125	0.2038	0.6096	0.1865	0.0000	0.8587
LAU1255	0.1807	0.2556	0.5637	0.0000	0.1505
LAU1314	0.8691	0.1055	0.0202	0.0005	0.0656
LAU335	0.6152	0.3626	0.0222	0.0000	0.0132
CIBERSORT (LM22)	B-cells	CD4+ T	CD8+ T	NK	SSQ
LAU125	0.6502	0.2631	0.0074	0.0793	0.1226
LAU1255	0.1659	0.2391	0.5576	0.0374	0.1555
LAU1314	0.6511	0.2350	0.1138	0.0000	0.0034
LAU335	0.6039	0.3181	0.0781	0.0000	0.0095
CIBERSORT (TRef)	B-cells	CD4+ T	CD8+ T	NK	SSQ
LAU125	0.5241	0.4453	0.0185	0.0121	0.3166
LAU1255	0.2357	0.2467	0.5176	0.0000	0.0997
LAU1314	0.7820	0.1713	0.0445	0.0022	0.0209
LAU335	0.7634	0.1814	0.0553	0.0000	0.0951
ICeD-T (pN, wN)	B-cells	CD4+ T	CD8+ T	NK	SSQ
LAU125	0.1668	0.7234	0.0930	0.0168	1.0316
LAU1255	0.2472	0.4517	0.2943	0.0068	0.1002
LAU1314	0.4998	0.3492	0.1367	0.0142	0.0422
LAU335	0.4824	0.4045	0.0983	0.0149	0.0030
ICeD-T (pN, wY)	B-cells	CD4+ T	CD8+ T	NK	SSQ
LAU125	0.1576	0.7438	0.0889	0.0096	1.0732
LAU1255	0.2099	0.0492	0.2830	0.0152	0.1381
LAU1314	0.6143	0.2944	0.0857	0.0056	0.0047
LAU335	0.5685	0.3731	0.0502	0.0081	0.0037
ICeD-T (pY, wN)	B-cells	CD4+ T	CD8+ T	NK	SSQ
LAU125	0.2162	0.6342	0.1188	0.0308	0.8508
LAU1255	0.2840	0.4282	0.2265	0.0613	0.0831
LAU1314	0.5068	0.3467	0.1337	0.0128	0.0389
LAU335	0.4530	0.4182	0.1063	0.0226	0.0078
ICeD-T (pY, wY)	B-cells	CD4+ T	CD8+ T	NK	SSQ
LAU125	0.1880	0.5705	0.2122	0.0293	0.8471
LAU1255	0.2806	0.5060	0.1551	0.0583	0.1381
LAU1314	0.5544	0.3223	0.1146	0.0087	0.0197
LAU335	0.5538	0.3708	0.0608	0.0147	0.0020

Table 10: Melanoma Data - Estimated relative proportions of Immune cells

It is clear from the above that CIBERSORT would produce the minimum sum of squared error

among all model fits due in chief to the manner in which it handles subject LAU125. ICeD-T with use of Tumor information (both with weight and no weight), produced the second best fit by sum of squared error. EPIC would produce the third best fit by sum of squares. Finally, ICeD-T without tumor purity would produce the worst results.

Examining subject LAU125, this subject is highly anomalous. Immune response in this subject is composed almost entirely of B-cells. Both EPIC and ICeD-T struggle to estimate the B-cell proportions for this subject - a likely consequence of their use of the same reference matrix. CIBERSORT does not struggle as greatly with this single subject and thus experiences smaller sums of squared error.

Across the remaining individuals, ICeD-T using any option produces the best results for LAU1255 and LAU335. ICeD-T without tumor purity and using maximal variance weights produces the best results for LAU1255, LAU1314 and LAU 335. Thus, outside of the subject LAU125, ICeD-T is able to provide the most competitive results across remaining subjects.

Focus now on the estimation of CD8 T-cell abundance across all methods. The use of ICeD-T without tumor purity provides the best fit across the singular cell type among all individuals.

E PD-1 Checkpoint Therapy Use in Melanomas

The final validation datasets for the ICeD-T method examine its application to a set of RNA-seq experiments derived from patients on PD-1 Checkpoint inhibitor therapies [5].

E.1 Data

The raw fastq files of RNA-seq data were downloaded from Sequence Read Archive (`https://www.ncbi.nlm.nih.gov/sra`), under the accession numbers SRP067938 and SRP090294. We mapped the RNA-seq reads to hg38 reference genome using STAR with gene annotation from GENCODE version 27. Then the number of RNA-seq fragments per gene were counted using R function `GenomicAlignments/summarizeOverlaps`.

E.2 Fit Method

CIBERSORT:

The CIBERSORT web application (Version: CIBERSORT Jar 1.06) was used to fit these TPM normalized RNA-seq data. The model was fit using the LM22 signature matrix and run with quantile normalization disabled.

EPIC:

The EPIC model is fit to the TRef reference matrix using TPM-normalized RNA-seq data.

ICeD-T:

The ICeD-T model is fit to the TRef reference matrix using TPM-normalized RNA-seq data. It is fit both without weights and with maximal variance weights derived from the TRef reference data. This is made possible through a function, `EPIC.Extract`, which extracts the fitted data and reference matrix from the EPIC library's function and outputs them in a form usable by ICeD-T.

TPM-normalization:

As noted above, data were provided in gene count form. As such, computation of TPM values using software such as RSEM is not possible. Thus, we transform the raw counts into TPM values using the following formula:

$$TPM_j = 10^6 \left(\frac{r_j/l_j}{\sum_{j=1}^{n_G} r_j/l_j} \right).$$

E.3 Additional Results

We examine the estimated probability being consistent for each of 98 genes in 28 samples. Each gene could be aberrant in one sample, but consistent in the other sample, and thus we cannot classify genes into these two classes (Figure 12). It does appear that in this dataset it is challenging to clearly assign a gene to be consistent or aberrant. Nevertheless, if we classify the entries of our data matrix of 98 genes times 28 samples into three groups based the probability being consistent (using the cutoffs of 33rd percentile and 67th percentile), and then draw scatter plots of model fit vs. observed expression, we can see that the deconvolution model fits the data much better when the probability of being consistent is high (Figure 13).

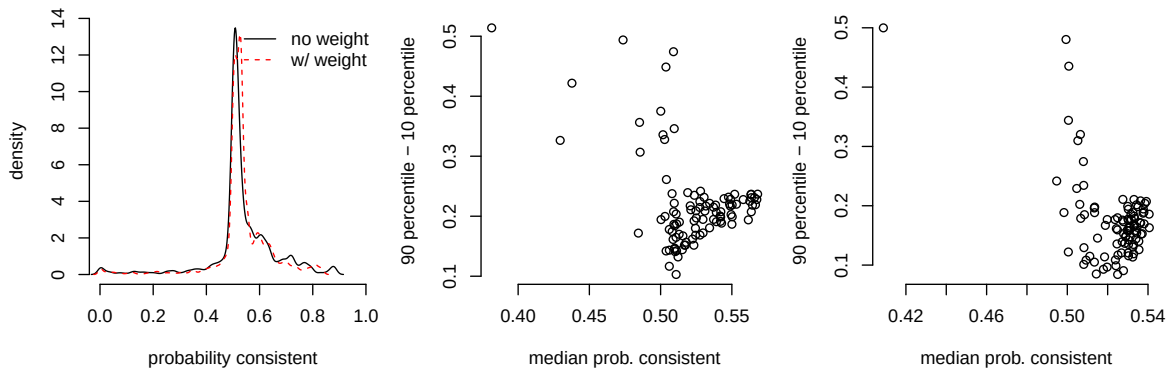


Figure 12: Left panel: distribution of consistent probability. Middle panel: summary of probability of being consistent for each gene across 28 samples by median (x-axis) and range (y-axis, 90 percentile - 10 percentile) without using weight. Right panel: similar to Middle panel, but ICeD-T model is fit using weight.

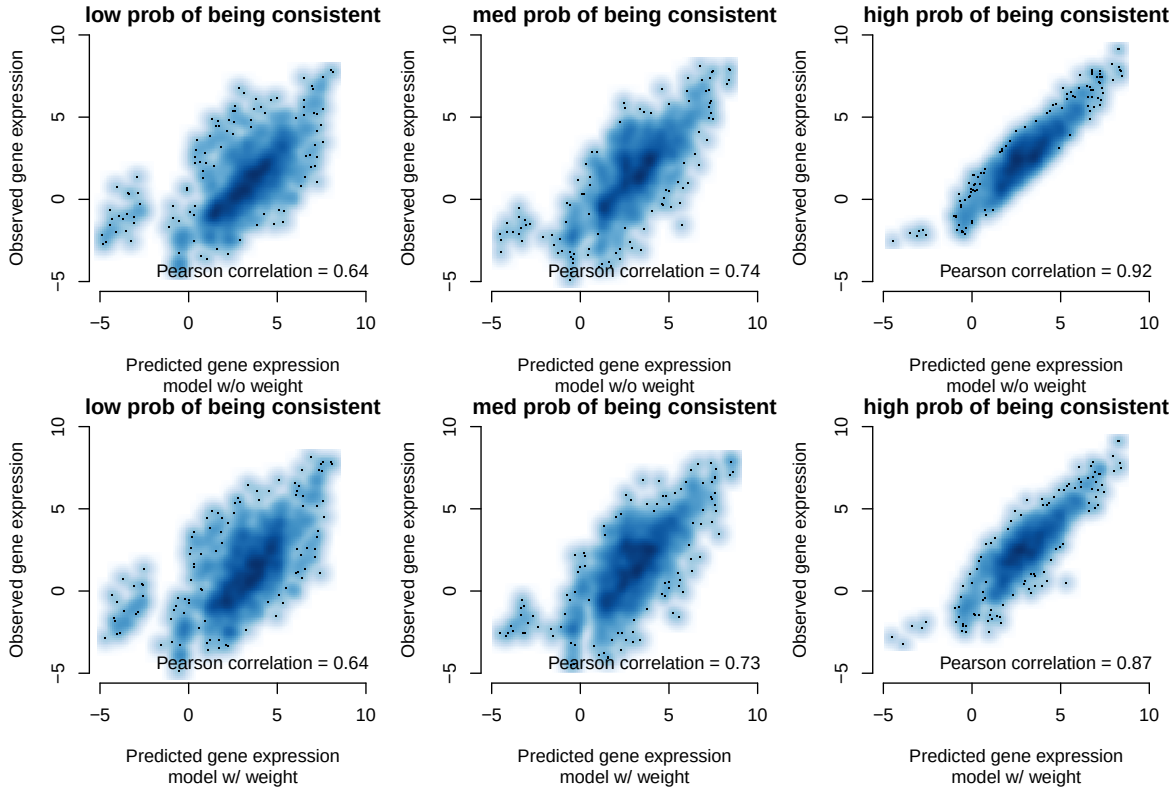


Figure 13: Examination of model fit if we divide the gene expression data matrix into three groups: with low, medium, or high probability of being consistent.

References

[1] 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. *Nature methods,* **12**(5), 453–457.

[2] Racle, J., deJonge, 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**, e26476.

[3] Linsley, P. S., 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.

[4] Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014) voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology,* **15**(2), R29.

[5] Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., 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.