HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2024 July 12.

Published in final edited form as:

Biometrics. 2023 June ; 79(2): 854–865. doi:10.1111/biom.13614.

Estimating cell type composition using isoform expression one gene at a time

Hillary M. Heiling[1], **Douglas R. Wilson**[1], **Naim U. Rashid**[1,2], **Wei Sun**[3], **Joseph G. Ibrahim**[1,2]

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

[2]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

[3]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington

Abstract

Human tissue samples are often mixtures of heterogeneous cell types, which can confound the analyses of gene expression data derived from such tissues. The cell type composition of a tissue sample may itself be of interest and is needed for proper analysis of differential gene expression. A variety of computational methods have been developed to estimate cell type proportions using gene-level expression data. However, RNA isoforms can also be differentially expressed across cell types, and isoform-level expression could be equally or more informative for determining cell type origin than gene-level expression. We propose a new computational method, IsoDeconvMM, which estimates cell type fractions using isoform-level gene expression data. A novel and useful feature of IsoDeconvMM is that it can estimate cell type proportions using only a single gene, though in practice we recommend aggregating estimates of a few dozen genes to obtain more accurate results. We demonstrate the performance of IsoDeconvMM using a unique data set with cell type–specific RNA-seq data across more than 135 individuals. This data set allows us to evaluate different methods given the biological variation of cell type–specific gene expression data across individuals. We further complement this analysis with additional simulations.

Keywords

alternative splicing; bulk expression; deconvolution; isoform; RNA-seq

Correspondence Hillary M. Heiling, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC. hheiling@live.unc.edu.

SUPPORTING INFORMATION

Web Appendix A, referenced in Section 2, and Web Appendix B, referenced in Sections 3, 4, and 5 are available with this paper at the Biometrics website on Wiley Online Library. An R package for the IsoDeconvMM method is available in gitHub: https://github.com/hheiling/IsoDeconvMM. The code used to run the analyses in the paper and Web Appendix B are provided both on the Biometrics website and GitHub: https://github.com/hheiling/IsoDeconvMM_Supplement

1 | INTRODUCTION

RNA sequencing data derived from human tissue samples are often mixtures of heterogeneous cell types. It is often of interest to quantify the relative abundance of each constituent cell type found within a tissue sample. In some cases, the relative abundance profiles themselves contain relevant information for the main goal of a study. For example, the relative abundance of different types of immune cells within tumor samples can be used to predict patients' response to cancer immunotherapy. In others, abundance profiles are crucial for proper cell type–specific differential expression analyses (Li and Wu, 2019; Jin et al., 2020). Cell-sorting and other physical separation techniques exist to partition tissue samples into purified samples of their constituent cell populations, but such methods can be costly and may even induce changes to the cellular environment, which can impact expression profiles (Shen-Orr et al., 2010). As an alternative to physical separation methods, the development of statistical models for the deconvolution of expression profiles from tissue samples has become an active area of research.

In silico expression deconvolution models can largely be separated into three main developments: ratio-based models, linear models, and infiltration scores. Ratio-based models rely upon computing expression ratios between a mixed expression profile and a "gold standard" reference for a single cell type. The minimum of these ratios across genes roughly approximates the proportion of the referent cell type (Gosink et al., 2007; Clarke et al., 2010; Wang et al., 2014). These methods are often limited to study two cell types (eg, tumor vs normal). The linear model and infiltration score approaches can handle more than two cell types. The linear model framework assumes that appropriately normalized mixture expressions can be modeled as a weighted summation of cell type–specific gene expression in two or more cell types (Lu et al., 2003; Gong and Szustakowski, 2013; Newman et al., 2015; Zhong et al., 2013). The infiltration scores approach aim to estimate unitless quantities designed to reflect the abundance each constituent cell type (Becht et al. (2016); Li et al. (2016)).

However, existing methods have been designed to utilize gene-level expression only. Thus, appropriate deconvolution requires that cell types express differently at the gene level. In the case of highly similar cell types, however, it may be the case that gene-level expression differences are minimal. An alternative is to quantify gene expression at a more granular level: isoform expression. Each gene in the human genome is often composed of multiple exons separated by introns, and one gene may produce multiple distinct transcripts by taking different combinations of exons. This process, known as alternative splicing, allows a single gene to encode multiple proteins and thus greatly increases the biodiversity of proteins that can be encoded by the genome. More than 90% of human genes could undergo alternative splicing (Wang et al., 2008). Because cell types are often defined through the expression of proteins, the isoform-level expression could be more sensitive to cell type identity than higher level gene expression that is often the summation of gene expression across multiple isoforms.

In this paper, we outline the development of a statistical model named IsoDeconvMM for expression deconvolution in mixture tissues by exploiting isoform-level expression

differences between cell types. A crucial factor for the success of expression deconvolution is to identify a good set of signature genes/isoforms whose expression has much higher variation across cell types than within cell types. However, even for such carefully selected genes/isoforms, there are still biological variation of cell type–specific gene/isoform expression across individuals. IsoDeconvMM is designed to explicitly model biological variability to achieve robust performance. We demonstrate the utility of our method using the Blueprint data set (Chen et al., 2016). This data set contains human bulk RNA-seq samples for three sorted immune cell populations (CD4-positive, alpha-beta T cell; CD14-positive, CD16-negative classical monocyte; and mature neutrophil) from up to 197 individuals. In an in silico data analysis, we used these data to model the variability across individuals and test the performance of our method given this biological variability.

The rest of the paper is organized as follows. In Section 2, we present the statistical models and algorithm used to estimate cell type proportions in mixture tissues, and describe the data and materials needed for the method. In Section 3, we present in silico data analyses. In these analyses, we compare the performance of our method with the performance of CIBERSORTx (Newman et al., 2015). In Section 4, simulations are conducted to assess the performance of the IsoDeconvMM method when different underlying data distributions are assumed. Concluding remarks are given in Section 5. Technical proofs are given in the Web Appendix A. Additional details regarding the procedures and materials required for the analyses in Sections 3 and 4 are given in the Appendix and the Web Appendix B.

2 | METHODS

2.1 | Required data and resources

Consider a biological tissue sample composed of K different cell types. We seek to estimate the unknown relative abundance of each cell type k—or the proportion of cells of type k—in the heterogeneous sample. In order to estimate these proportions, IsoDeconvMM requires a single RNA-seq experiment performed on the mixture sample. In addition, it is assumed that there exist RNA-seq data for N_k purified samples for cell type k. For each sample, RNA-seq read counts are summarized at the exon level by counting the number of reads (or RNA-seq fragments for paired-end reads) overlapping various sets of exons.

In order to summarize the read counts at the exon set level, it is assumed that a detailed gene model on the location of each exon and the structure of each isoform is available for each gene. Consider a hypothetical gene composed of m nonoverlapping exons that are utilized by I isoforms, or distinct mRNA transcripts formed by unique combinations of these exons. As specified in the gene models, the locations of these exons within the gene are known as are the identities and compositions of all isoforms used by this gene. We define the read count at any exon set A as the number of reads that overlap each of the exons in A and only these exons.

To visualize the setup, consider the hypothetical gene displayed in Figure 1. This gene is composed of $E = 4$ exons. An exon set is defined as some subset of the exons, which for this hypothetical gene could include sets containing only a single exon, sets containing two of the four exons, sets containing three of the four exons, or the set with all four exons

combined. Each RNA-seq read from the gene maps to one and only one of the possible exon sets. If an RNA-seq read maps to each exon in some exon set A and no other exons, we say it belongs to exon set A.

The gene in Figure 1 is composed if $I = 3$ isoforms. Suppose that isoforms 1, 2, and 3 compose the set of all isoforms used by the gene and that their structure with respect to the exons is as given in the figure. Consider the exon set $A := \{1, 2, 3\}$. The read count at A is defined as the number of RNA-seq reads, which, when mapped, overlap exons 1, 2, and 3 but do not overlap exon 4.

Identifying the exon set to which an RNA-seq read belongs gives us insight into the isoform to which the read belongs. Although a gene is composed of $(2^E - 1)$ possible exon sets, the exon sets possible for each of the isoforms can be restricted. In this hypothetical example, isoforms 1 and 2 do not contain exon 3, so none of the exon sets containing exon 3 are possible for isoforms 1 and 2. Which exon sets are theoretically possible for each of the three isoforms of this gene is provided in Table 1.

In some cases, two exons of a gene overlap partially. When this happens, we handle the situation similar to Sun et al. (2015). We split the two exons into three exons: the two nonoverlapping sections unique to a particular exon and the overlapping section belonging to both exons. It is also possible for multiple genes to overlap one or more exons, and we consider these overlapping genes as a transcript cluster.

IsoDeconvMM also assumes that there exists a list of cell type–specific genes wherein there are gene- and/or isoform-expression differences across the K cell types. Such a list of genes can be found using one of a variety of expression testing methods for RNA-seq data. Furthermore, IsoDeconvMM requires empirical knowledge of the fragment length distribution for the bulk RNA-seq samples.

2.2 | The IsoDeconvMM model and algorithm

2.2.1 | Model parameters—

Within the IsoDeconvMM model, cell type proportions are estimated independently within each gene, and these gene-specific proportion estimates are then aggregated to produce a sample-level cell type relative abundance estimate. To simplify discussion, we outline the IsoDecon-vMM model for a single gene.

RNA-seq expression is commonly corrected for feature length. Previously, however, the notion of feature length pertained to the length of the genes or isoforms being measured and not to the lengths of exon sets. Sun et al. (2015) extended the definition of feature length to exon sets and referred to it as the effective length for exon sets. Briefly, the effective length of an exon set is the expected number of starting locations, where an RNA-seq fragment that overlaps with all the exons of this exon set can be sampled. Such expectation is taken over the distribution of RNA-seq fragment length. Note that the effective length of an exon set varies across isoforms. For example, isoforms that do not contain all the exons within the set cannot produce reads in that exon set, thus the effective length of the exon set for such isoforms will be zero. See the supplementary material of Sun et al. (2015) for more details.

We first consider the models and parameters used to describe the gene expression in cell type–specific samples. In all the notation, we utilize the subscripts kj to denote the parameters for sample j of cell type K. Let $\boldsymbol{Y}_{kj} = \{Y_{kjA}\}$ denote the vector of read counts across all E exon sets in the given gene/transcript cluster for sample j of cell type K. Also denote $Y_{kj(O)}$ as the total read count outside the gene of interest in this sample. We assume that the vector $\left(Y_{kj(O)}, Y_{kj}^T\right)^T$ follows a multinomial distribution

$$\begin{bmatrix} Y_{k\,j(O)} \\ \boldsymbol{Y}_{kj} \end{bmatrix} \Big| \tau_{kj}, \boldsymbol{\gamma}_{kj} \sim \text{Multinomial}\left(t_{kj}, \begin{bmatrix} 1-\tau_{kj} \\ \tau_{kj}\boldsymbol{X}\boldsymbol{\gamma}_{kj} \end{bmatrix}\right), \tag{1}$$

where τ_{kj} is the probability that a randomly selected read maps to the gene of interest, $\boldsymbol{\gamma}_{kj} = (\gamma_{kj1}, \ldots, \gamma_{kjI})^T$ is the vector of I isoform expression parameters, t_{kj} is the total read count in the sample, and $\boldsymbol{X}$ is a matrix of effective lengths such that column i of $\boldsymbol{X}$ is the vector of effective lengths for all the exon sets of isoform i.

We further describe the probability τ_{kj} and the isoform parameters $\boldsymbol{\gamma}_{kj}$ with the following beta and Dirichlet distributions:

$$\begin{aligned} \tau_{kj} &\sim \text{Beta}(\boldsymbol{\beta}_k) \\ \tilde{\boldsymbol{l}} \circ \boldsymbol{\gamma}_{kj} &\sim \text{Dirichlet}(\boldsymbol{\alpha}_k), \end{aligned} \tag{2}$$

where $\tilde{\boldsymbol{l}} = \left(\tilde{l}_1, \ldots, \tilde{l}_I\right)$, $\tilde{l}_i = \sum_{A \in \text{isoform } i} \boldsymbol{X}_A$ represents the total effective lengths of isoform i for $1 \le i \le I$, and $\circ$ represents element-wise multiplication of two vectors. It should be noted that the $\boldsymbol{\gamma}_{kj}$ parameters can be interpreted as per-unit-of-effective-length conditional probabilities that a read maps to isoform i given that it maps to the gene, which utilizes isoform i. The fact that we model gene expression for each sample j of cell type k separately in the above models allows us to capture the biological variation across samples. The similarity of all the samples from cell type k is modeled by the shared beta or Dirichlet distribution in Equation (2). We next consider the models and parameters used to describe the exon set counts in the mixture sample. Let $\boldsymbol{Z} = \{Z_A\}$ denote the vector of read counts across all E exon sets in the given gene for the mixture sample, and let $Z_T = \sum_A Z_A$ denote the sum of the read counts for the given gene. We assume that the vector of counts $\boldsymbol{Z}$ follows a multinomial distribution such that

$$[\boldsymbol{Z}] \mid \tau_k^*, \boldsymbol{\gamma}_k^* \sim \text{Multinomial}\left(Z_T, \left[\frac{\sum_{k=1}^K \rho_k \tau_k^* \boldsymbol{X}\gamma_k^*}{\sum_{k=1}^K \rho_k \tau_k^*}\right]\right), \tag{3}$$

where τ_k^* represents the probability that a randomly selected read from cell type k maps to the gene of interest in the mixture sample, $\boldsymbol{\gamma}_k^* = (\gamma_{k1}^*, \ldots, \gamma_{kI}^*)$ is the vector of I isoform

expression parameters unique to cells of type k found within the mixture sample, and ρ_k is the proportion of cell type k in the mixture sample.

Using the same cell type k gene expression hyperparameters $\boldsymbol{\beta}_k$ and isoform expression hyperparameters $\boldsymbol{\alpha}_k$ from the pure sample models in Equation (2), we further describe the probabilities τ_k^* and the mixture isoform parameters $\boldsymbol{\gamma}_k^*$ as follows:

$$\begin{aligned} \tau_k^* &\sim \text{Beta}(\boldsymbol{\beta}_k) \\ \tilde{\boldsymbol{l}} \circ \boldsymbol{\gamma}_k^* &\sim \text{Dirichlet}(\boldsymbol{\alpha}_k)\,. \end{aligned} \tag{4}$$

Given those shared parameters $\boldsymbol{\beta}_k$ and $\boldsymbol{\alpha}_k$, we assume independence across samples.

2.2.2 | Model estimation—Within each gene, the model is fit using a staged estimation approach with three stages. In Stage 1, the gene and isoform expression parameters are estimated separately for each purified reference sample by maximum likelihood estimation. The likelihood used for Stage 1 involves only Equation (1). Under such a framework, closed-form estimates of τ_{kj} are obvious and a logarithmic adaptive barrier algorithm can be used to obtain estimates of the $\boldsymbol{\gamma}_{kj}$ subject to boundary constraints. Once obtained for each cell type and sample, these estimates are held fixed for all further stages.

Within Stage 2, the values of τ_{kj} and $\boldsymbol{\gamma}_{kj}$ estimated during Stage 1 are treated as observations from Equation (2). Estimates of $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ are obtained via maximum likelihood estimation within separate Dirichlet models. Once obtained, these estimates of $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ are fixed for Stage 3.

Finally, in Stage 3, the $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ estimates are used in Dirichlet distributions as penalty functions in the estimation of the $\boldsymbol{\gamma}_k^*$ and τ_k^*. In this way, we regularize estimates of $\boldsymbol{\gamma}_k^*$ and τ_k^* to be similar to those estimates obtained in the pure cell type samples. The use of an Expectation-Maximization (EM) algorithm allows separation of the full likelihood into $K+1$ independent components in the M step. The first K components pertain to the isoform expression parameters from each of the K cell types. Each of these components is optimized using a Newton-Raphson algorithm on the $\log(\boldsymbol{\gamma}_k^*)$ until convergence of isoform parameters. The last component contains information regarding the ρ_k and $\log(\tau_k^*)$ values, which are optimized using a quasi-Newton's method optimization procedure (Broyden-Fletcher-Goldfarb-Shanno). Estimation is seeded at various start points to identify global maxima. The E step updates the posterior means of the exon set counts in the mixture sample ($\boldsymbol{Z}$) attributable to cell type k. The expectation has a closed-form solution, provided in Web Appendix A. The EM algorithm is iterated until convergence in the proportion estimates. Proportion estimates across multiple genes are then aggregated using the spatial median to obtain final estimates of cell type proportions.

Technical proofs and further details about the models and methods can be found in Web Appendix A located in the online supplementary information. Web Table 1 in Web Appendix

A contains a summary of the notation presented in Section 2. A discussion about why a staged estimation approach was used instead of a joint estimation approach is included in Section Web Appendix A1.7.

3 | IN SILICO BLUEPRINT DATA ANALYSIS

To the best of our knowledge, our IsoDeconvMM is the first method that estimates cell type proportions using isoform expression. Since there are already several methods for cell type composition estimation using gene expression (instead of isoform expression) data, an immediate question is what is the advantage to use isoform expression. In this section, we compare our IsoDeconvMM method with CIBERSORTx (Newman et al., 2019), a representative and popular method for cell type composition estimation using gene expression, and demonstrate that IsoDeconvMM has similar performance with CIBERSORTx when the number of genes is relatively large and it outperforms CIBERSORTx with large margin when the number of genes is small. To compare IsoDeconvMM and CIBERSORTx, we utilize the Blueprint data set (Chen et al., 2016) discussed in Section 1. We arbitrarily label the three cell types as follows: CT1 represents CD4-positive, alpha-beta T cell; CT2 represents CD14-positive, CD16-negative classical monocyte; and CT3 represents mature neutrophil.

In order to create mixture files from the Blueprint data, we selected 100 individuals who had pure reference samples collected from all three cell types. For each of these individuals, we used their pure reference samples to create a mixture file. The 100 mixture proportions were randomly selected from the distribution $\rho_{mix} \sim$ Dirichlet(2, 2, 2). Relatively extreme probabilities, defined as probability vectors that assigned one or more cell types to have a probability less than .05, were eliminated from consideration.

To select genes/transcript clusters to be used by the IsoDeconvMM method, we sought to identify differential isoform usage (DU) transcript clusters that had the largest difference between the isoform distributions in the three cell types. To this end, we identified clusters that had at least one isoform highly expressed in one cell type and either minimally expressed or not expressed at all in the other two cell types, collectively. The selection of transcript clusters proceeded as follows. We selected 10 pure reference samples (not used in the mixture file creation) from each of the three cell types present in the Blueprint data. We then used the isoDetector function in the isoform R package (Sun et al., 2015) to acquire isoform abundance information for transcript clusters present on chromosomes one through four for all of the 30 pure reference samples. Using the abundance information output, we examined both fold change magnitudes and Wilcoxon rank sum tests comparing abundance levels for the isoforms in the cluster between a single cell type and the other two cell types combined. Using these results, we identified isoforms of interest. The transcript clusters that these isoforms belonged to were then selected for further analysis. A full description of the procedure to identify DU clusters of interest can be found in the Appendix.

For the CIBERSORTx method, we aimed to select DE transcript clusters in a similar manner to the DU transcript clusters used in the IsoDeconvMM analysis. We first quantified the total expression per transcript cluster, restricting the transcript clusters considered to those present

on chromosomes one though four. Then we applied DESeq2 (Love et al., 2014) to identify transcript clusters with differential expression that were relatively overexpressed in one cell type compared to the other two cell types combined.

In the Web Appendix B, we compared the performance of the IsoDeconvMM algorithm across different algorithm settings. Based on results presented in the Web Appendix B (see Web Figure 1), we concluded that using five samples per cell type in the IsoDeconvMM analysis was sufficient. Therefore, all further IsoDeconvMM and CIBERSORTx results utilize five pure reference samples per cell type. Since the IsoDeconvMM algorithm requires multiple initial points in order to optimize the accuracy of the results, we also explored how many initial points were sufficient to use. Web Figure 2 in the Web Appendix B suggests that using the 10 generic initial points specified in Table A1 in the Appendix is sufficient for this case of three total cell types. Therefore, all IsoDeconvMM algorithm results presented in this section utilized these 10 initial points in the algorithm. Recommendations of initial points for the generic case of K cell types are given in the Appendix. The IsoDeconvMM package gives automated recommendations for initial points.

In the exploratory analyses presented in the Web Appendix B, we found that the estimates of the cell type–specific isoform parameters could be unstable for a small number of transcript clusters. This could be due to extra variance or outliers in these genes. In those clusters, the estimate of the α_k parameters of Equation (2) (estimated in Stage 1 of the model fit algorithm) tended be much larger than other clusters. Therefore, we performed a filtering step such that if two or more cell type–specific isoform parameter estimates for a transcript cluster were greater than 500, the cluster was excluded from further analysis. We now compare the the performance of IsoDeconvMM and CIBERSORTx results when different numbers of transcript clusters were used in the analysis (Figure 2). In each simulation setup, the best N of the available transcript clusters were selected by first choosing the best n_s clusters per cell type comparison (cell type j vs the other two cell types collectively) and then take their union. The number n_s was adjusted such that the union gave $N = \{100, 50, 25, 10\}$ clusters.

When 100 or 50 transcript clusters are used in the analysis, both the CIBERSORTx and IsoDeconvMM methods perform well, with CIBERSORTx performing slightly better than IsoDeconvMM. For the 25 cluster case, both methods perform equally well. For the case when only 10 clusters are used, the CIBERSORTx method is very unstable. In contrast, the IsoDeconvMM method is still reasonably accurate.

4 | SIMULATION STUDIES

Our model assumes an underlying Dirichlet-multinomial distribution, which allows overdispersion beyond the variance of multinomial distribution. However, it is still possible that a Dirichlet-multinomial distribution cannot fit the real data well. In this section, we evaluate the performance of IsoDeconvMM when the observed data are generated from Dirichlet-negative binomial distributions. We simulated bulk RNA-seq counts data from three sorted cell populations given the generic labels of CT1, CT2, and CT3.

For all of the simulations, we first generated the gene-level counts from a Dirichlet-multinomial distribution. In order to make the distribution of gene counts as realistic as possible, we used gene count distribution from a real data set (Parikshak et al., 2016) that contained the number of RNA-seq reads per gene for 89 human bulk RNA-seq samples. The genes present in this data set were filtered such that each transcript cluster was comprised of a single gene (for convenience purposes) and genes with low expression were excluded. The genes were then limited to those present on chromosomes one to nine in order to reduce computational burden, resulting in 5172 total genes. A full description of the gene selection procedure is provided in the Web Appendix B. We fit a Dirichlet distribution to these data using the R package DirichletReg (Maier, 2014).

For each simulated pure sample, the Dirichlet distribution described above generated a probability vector associated with the genes. The total read count per sample was selected from a normal distribution with mean 7 million and standard deviation 1 million. This normal distribution was based on the distribution of the total read counts of the selected 5172 genes in the 89 bulk RNA-seq samples (Parikshak et al., 2016). Individual gene counts were then generated using a multinomial distribution.

Of the total 5172 genes, we selected 1000 genes with relatively high expression and at least three isoforms as possible genes to be used for the mixture sample proportion estimate in the IsoDeconvMM analysis. For all 1000 of these genes of interest, we calculated the effective length design matrix $\boldsymbol{X}$ as described in Section 2. After additional filtering to exclude genes with over 15 isoforms, we randomly selected 100 genes for DU.

The Dirichlet-multinomial and the Dirichlet-negative binomial simulations diverge on the simulation of the exon set counts. For each cell type, we gave each of the 1000 genes of interest a Dirichlet distribution for their isoforms. These Dirichlet distributions only differed between the three cell types for the 100 genes specified for DU. A probability vector $\boldsymbol{\pi}_g = (\pi_{g1}, \ldots, \pi_{gI})$ was drawn from these Dirichlet distributions, where π_{gi} is the proportion of read counts in isoform i given that the read comes from gene g.

In the Dirichlet-multinomial simulation, the vector π_g was set equal to the $\tilde{l} \circ \gamma_g$ vector described in Section 2 in Equation (2). We then model the exon set counts for gene g by

$$\boldsymbol{y}_g \sim \text{Multinomial}(T_g, s_g \sum_{i=1}^{I} \boldsymbol{x}_{gi}\gamma_{gi}), \quad \gamma_{gi} \geq 0, \tag{5}$$

where $\boldsymbol{x}_{gi}$ for $1 \leq i \leq p$ represents the vector of effective lengths of all of the exon sets for the ith isoform for gene g, T_g is the total read count for gene $s_g \sum_{i=1}^{I} \boldsymbol{x}_{gi}\gamma_{gi} = 1$, and S_g is the scaling factor such that $s_g \sum_{i=1}^{I} \boldsymbol{x}_{gi}\gamma_{gi} = 1$.

In the Dirichlet-negative binomial simulation, the probability vector π_{gi} was used differently. The vector of counts of the possible exon sets within the gene, $\boldsymbol{y}_g$, was given a negative binomial distribution $\Psi(\boldsymbol{\mu}_g, \boldsymbol{\phi})$ with mean μ_g and dispersion parameter $\boldsymbol{\phi}$. We model $\boldsymbol{\mu}_g$ by

$$\boldsymbol{\mu}_g = \boldsymbol{X}_g\boldsymbol{\beta}_g = \sum_{i=1}^{p} \boldsymbol{x}_{gi}\beta_{gi} = \sum_{i=1}^{p} \boldsymbol{x}_{gi}\pi_{gi}r_g, \quad \beta_{gi} \geq 0 \tag{6}$$

where π_{gi} again is the proportion of read counts in isoform i given that the read comes from gene g, $\boldsymbol{X}_g = (\boldsymbol{x}_{g1}, \ldots, \boldsymbol{x}_{gp})$, $\boldsymbol{x}_{gi}$ for $1 \leq i \leq p$ represents the effective lengths of all of the exon sets for the ith isoform for gene g, and r_g is a scaling factor equal to the ratio of the total read count of the gene and the sum of the vector $\sum_{i=1}^{p} \boldsymbol{x}_{gi}\pi_{gi}$.

In the Dirichlet-negative binomial simulations, we also compared the algorithm fit results under low and moderate overdispersion assumptions for the Negative Binomial portion of the model. The dispersion parameter ϕ was given the range 1/90 to 1/120 for the low dispersion setup and the range 1/50 to 1/60 for the moderate dispersion setup.

In order to make the isoform Dirichlet distributions for the DU genes as realistic as possible, we modeled these distributions using the results from the Blueprint data set analysis described in Section 3. The cell type–specific isoform Dirichlet parameter $\boldsymbol{\alpha}_k$ (estimated by Dirichlet-multinomial distribution) were used in the simulations. In the Dirichlet-multinomial simulations, these values were used directly. In the Dirichlet-negative binomial simulations, these values were multiplied by a constant of five so that the overall variance between Dirichlet-multinomial and Dirichlet-negative binomial are similar.

Three data sets were simulated using the three different data modeling assumptions: Dirichlet-multinomial, Dirichlet-negative binomial with moderate overdispersion, and Dirichlet-negative binomial with low overdispersion. We generated 15 pure reference samples per cell type for each simulation setup. We partitioned the pure samples such that for each cell type, 10 samples were used to generate the mixture samples and the other 5 were used to estimate cell type–specific gene/isoform expression. Fifty mixture proportions were randomly selected from the distribution $\rho_{mix} \sim$ Dirichlet(2, 2, 2). Relatively extreme probabilities, defined as probability vectors that assigned one or more cell types to have a probability less than .05, were eliminated from consideration.

For the fragment length distribution file, we chose to simulate paired-end read lengths from a truncated normal distribution with mean 300 bp, standard deviation 50 bp, and truncated to the left at 150 bp. For the initial points, we used the same 10 generic initial points used in Section 3, provided in the Appendix in Table A1.

All three of the simulated data sets were then fit using the IsoDeconvMM algorithm and we examine the performance of IsoDeconvMM when different number of transcript clusters are used to estimate cell type proportions. In each simulate setup, we randomly selected the desired number of transcript clusters from the 100 simulated DU clusters. The results presented in Figures 3 and 4 suggest that the results of our IsoDeconvMM method is robust to the data generation mechanisms. The only situation where the performance of

IsoDeconvMM is slightly worse is when the number of transcript clusters is small (ie, only 10 clusters) and the Dirichlet-negative binomial has moderate overdispersion.

5 | DISCUSSION

We have developed a new statistical method named IsoDeconvMM that estimates cell type abundance of bulk RNA-seq samples that are mixtures of multiple cell types. This method is unique from other deconvolution methods in that it utilizes DU information. We anticipate that this method will be of particular relevance in cases where DU is more informative than differential gene expression, or when the number of available genes is small. Currently, application of our method is limited by the availability of cell type–specific and isoform-specific gene expression data. Single cell RNA-seq (scRNA-seq) is a popular approach to generate cell type–specific gene expression data across different cell types, though most scRNA-seq pipelines cannot capture the complete information of different isoforms. However, the emerging spatial RNA-seq data show that it is possible to capture isoform-level gene expression for each cell or a few cells around a locus (Lebrigand et al., 2020; Maynard et al., 2020). We expect that the full advantage of IsoDeconvMM can be demonstrated when combining such cell type–specific and isoform-specific expression derived from these new pipelines.

We did not know of another deconvolution method that utilizes DU information with which to compare our method. Instead, we compared IsoDeconvMM with CIBERSORTx (Newman et al., 2015), which utilizes differential gene expression information. We believe a key advantage of our method over existing reference-based deconvolution methods is that we can estimate cell type fractions using the gene expression data from a single gene by exploiting the relative expression of each isoform within a gene. We tested this theory by comparing our method with CIBERSORTx, which uses information across genes. We found that our method performs similarly compared with CIBERSORTx when a moderate number of genes or transcript clusters are used and outperforms CIBERSORTx when a small number of transcript clusters are used. In addition to seeing this pattern in the in silico Blueprint analyses presented in Section 3, we also found similar results when we performed both IsoDeconvMM and CIBERSORTx on simulated Dirichlet-multinomial data (see Web Appendix B for details and results). This could be very useful when it is desired to distinguish between highly similar cell types, such as closely related neuron cells, in which case there may not be many transcript clusters that can truly discriminate between the cell types. Additionally, this could be useful in clinical settings that utilize a small panel of genes.

Although we could have compared our method with CIBERSORTx using isoforms instead of genes (or transcript clusters), thereby using information across isoforms, we felt applying CIBERSORTx on isoforms has several limitations. The estimate of isoform expression is generally more noisy and has more measurement error. Furthermore, a major limitation of approaches that use information across genes/isoforms is that a sufficient sample size of genes or isoforms is required. This limitation, which was illustrated in the in silico analysis results, is the same limitation for either CIBERSORTx on genes or CIBERSORTx on isoforms. Consequently, using CIBERSORTx on isoforms would not provide a benefit. In

contrast to methods that use information across genes/isoforms, IsoDeconvMM utilizes gene expression variation across exon sets. The number of exon sets can increase quickly with the number of exons, and thus there are many genes with enough sample size within a gene itself.

The IsoDeconvMM method has other beneficial properties. In Web Appendix B, we have demonstrated that our method only requires a small number of pure reference samples per cell type. The simulations also show that the IsoDeconvMM method is robust to some model misspecification.

The IsoDeconvMM method has some limitations related to its computation time. Part of the reason for this time limitation is due to the fact that it requires an input of multiple initial points. However, this could be remedied using parallel computation techniques. Parallel computation techniques can be easily used in conjunction with the IsoDeconvMM method because a separate proportion estimate is calculated for each transcript cluster, and these individual estimates are later aggregated to get the overall proportion estimate. The IsoDeconvMM package, available for download in gitHub, allows for either serial or parallel computation. When the algorithm was run in serial using the UNC Longleaf computing cluster (CPU Intel processors between 2.3 GHz and 2.5 GHz), it took an average of 16.55 minutes to estimate the mixture proportion for a transcript cluster using 10 initial points.

More generally, the IsoDeconvMM procedure has the same limitations that apply to all reference-based deconvolution methods. These methods require assumptions about the true number and identity of cell types in the mixture samples. In many applications, this cell type information is unknown.

We looked further into the cell type 1 bias seen in the in silico Blueprint analyses. We performed 10 replicates of the in silico analyses, picking different sets of 100 individuals to create the mixture samples, picking different sets of pure reference samples, but using the same transcript clusters used in Section 3. We found that 2 of the 10 analysis replicates resulted in similar V shapes in the cell type 1 scatter plots seen in the paper results, but the other 8 replicates did not. This led us to believe that this concerning V shape in the cell type 1 scatter plot results were likely a result of unlucky randomness in the simulation setup. See Web Appendix B for further details and results.

It should be noted that the IsoDeconvMM method is sensitive to the isoform distribution effect size across the different cell types. We recommend users to be conscientious about selecting isoforms with the greatest effect sizes between the different cell types, regardless of what method they choose to identify isoforms with differential usage across the cell types.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding information

Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill

DATA AVAILABILITY STATEMENT

The Blueprint data that support the findings of this paper are available from BLUEPRINT Data Access Committee Members. Restrictions apply to the availability of these data, which were used under license for this paper. Data are available from the authors with the permission of BLUEPRINT Data Access Committee Members (blueprint-dac@ebi.ac.uk).

APPENDIX: ADDITIONAL SIMULATION DETAILS

A.1 | Transcript cluster selection: Large isoform effect sizes

This section describes the procedure used to select transcript clusters that have isoforms highly expressed in one cell type but minimally expressed or not expressed at all in the other two cell types. In the first part of the section, we describe the specific procedure we utilized to select transcript clusters for use in the in silico Blueprint analysis described in Section 3; many of the pure sample isoformparameter estimates from these clusters were also used in the simulations presented in Section 4. Later in the section, we discuss how these steps can be generalized for those wishing to use the IsoDeconvMM procedure.

A.1.1 | In silico Blueprint analysis

Ten samples per cell type were selected from the cell type–specific gene expression data generated by the Blueprint project (Chen et al., 2016). These 30 samples were separate from the samples used during the IsoDeconvMM algorithm fit and the samples used to create the mixture files. For each of these samples, the function isoDetector from the isoform R package (Sun et al., 2015) was applied to obtain penalized estimation of isoform-level expression for each cluster.

Next we outline the procedure for cluster selection. The transcript clusters were first filtered such that we only considered clusters on chromosomes one through four. Additionally, transcript clusters were filtered such that every cluster had between 3 and 20 isoforms.

For each cell type, we sought to select a cluster if it has at least one isoform with high expression in one cell type, and no or minimal expression in all other cell types. The same procedure is applied to each cell type and here we just use cell type 1 as an example. For each transcript cluster, we identified isoforms that were sufficiently expressed in cell type 1 (eg, it had nonzero abundance values in at least 9 of the 10 samples for cell type 1). For each isoform that met this criteria, we calculated the fold change of its average abundance in cell type 1 versus the average abundance in the other two cell types combined.

In addition to fold change, we also applied hypothesis testing for cluster selection. Again, consider cluster selection for cell type 1. We again identified isoforms that were sufficiently expressed in cell type 1 (eg, it had nonzero abundance values in at least 9 of the 10 samples for cell type 1). For each isoform that was expressed in cell type 1, a one-sided Wilcoxon rank sum test was performed to test the hypothesis that this isoform has higher abundance in cell type 1 than the other two cell types combined.

Isoforms that resulted in Bonferroni-adjusted P-values below the .05 threshold from the Wilcoxon rank sum tests were kept for further consideration. Of the isoforms that met this criteria, the 60 isoforms with the largest fold change values from each cell type were selected. The union across all cell types of the clusters associated with these best isoforms gave 130 transcript clusters.

Once the pure sample fit portion of the IsoDeconvMM algorithm was applied to these transcript clusters, some further filtration was applied. Clusters whose pure sample isoform Diriclet parameter values resulted in NA values or extremely large and divergent values (more than two values were greater than 500) were excluded from further consideration. In the case when five pure samples were used to estimate the cell type–specific parameters, eight clusters met this exclusion criteria.

Once these clusters were excluded, n_s isoforms with the greatest fold change values for each cell type were selected. We adjusted the value of n_s so that the total number of transcript clusters selected was 100, 50, 25, and 10.

A.1.2 | Generalization of procedure

We provide here a generalization of the above procedure. For general data with K cell types, we recommend obtaining at least five pure cell type reference samples from each cell type. On each of the pure cell type reference samples, run the isoDetector function in order to obtain the abundance estimates of each isoform within each transcript cluster. For a particular cell type k, perform the following steps:

1. Identify isoforms where no more than one of the pure reference samples for cell type k have an estimated abundance of zero for that isoform.
2. For each isoform that meets the criteria of step 1, calculate the average abundance of the isoform within the samples of cell type k and calculate the average abundance of the isoform within all other cell type samples. Calculate the fold change between these average estimates.
3. For each isoform that meets the criteria of step 1, perform a one-sided Wilcoxon rank sum test to test the hypothesis that this isoform has higher abundance in cell type 1 than the other two cell types combined. Calculate Bonferroni-adjusted P-values and ignore isoforms that give adjusted P-values above a certain cutoff (eg, cut-off .05).
4. Of the isoforms that meet the criteria of step 3, examine their fold change estimates. At this step, one could either pick the X isoforms with the highest fold change values (eg $X = 50$ or $X = 25$) or pick the isoforms with fold change values above a particular threshold.
5. For the isoforms picked after step 4, identify the transcript clusters to which these isoforms belong.

TABLE A1

The 10 generic initial points used in the in silico Blueprint analysis

CT1	CT2	CT3
0.10	0.10	0.80
0.10	0.80	0.10
0.80	0.10	0.10
0.25	0.25	0.50
0.25	0.50	0.25
0.50	0.25	0.25
0.20	0.40	0.40
0.40	0.20	0.40
0.40	0.40	0.20
0.33	0.33	0.33

Complete the above procedure for each cell type $k = 1, \ldots, K$. Use the transcript clusters identified with this procedure in the IsoDeconvMM analysis.

A.1.3 | Initial points used for in silico Blueprint analysis

Table A1 comprises a systematic approach to selecting initial points, where the following scenarios are represented: extreme cases where one cell type dominates with a large proportion and the other cell types split the remaining proportion; the equality case where all cell types are represented equally; and moderate cases that fall in between the extreme and equality cases. In the more general case of K cell types, we would also recommend setting up a mix of these three cases for the initial points. For the extreme cases, one could consider setting initial points in the following manner: $K - 1$ cell types initialized with proportion 0.10, and the Kth cell type initialized with the remaining proportion $(1 - 0.1 * (K - 1))$. When $K \geq 4$, it would be sufficient to leave out moderate cases and instead just add the equality case when each proportion is equal to $1/K$, which would not be much different from any moderate cases that could be specified. In the case of $K = 2$, we recommend adding the moderate cases of cell type proportion combinations $\{0.25, 0.75\}$ and $\{0.33, 0.67\}$. The IsoDeconvMM R package automatically recommends initial points in the above manner.

REFERENCES

Becht E, giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F. et al. (2016) Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biology, 17, 1–20. [PubMed: 26753840]

Chen L, ge B, Casale FP, Vasquez L, Kwan T, garrido-Martín D. et al. (2016) genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell, 167, 1398–1414. [PubMed: 27863251]

Clarke J, Seo P. and Clarke B. (2010) Statistical expression deconvolution from mixed tissue samples. Bioinformatics, 26, 1043–1049. [PubMed: 20202973]

Gong T. and Szustakowski JD (2013) Deconrnaseq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-seq data. Bioinformatics, 29, 1083–1085. [PubMed: 23428642]

Gosink MM, Petrie HT and Tsinoremas NF (2007) Electronically subtracting expression patterns from a mixed cell population. Bioinformatics, 23, 3328–3334. [PubMed: 17956877]

Jin C, Chen M, Lin D. and Sun W. (2020) Cell type aware analysis of RNA-seq data (carseq) reveals difference and similarities of the molecular mechanisms of schizophrenia and autism. bioRxiv.

Lebrigand K, Bergenstråhle J, Thrane K, Mollbrink A, Barbry P, Waldmann R. et al. (2020) The spatial landscape of gene expression isoforms in tissue sections. bioRxiv.

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

Li Z. and Wu H. (2019) Toast: improving reference-free cell composition estimation by cross-cell type differential analysis. Genome Biology, 20, 190. [PubMed: 31484546]

Love MI, Huber W. and Anders S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. Genome Biology, 15, 550. [PubMed: 25516281]

Lu P, Nakorchevskiy A. and Marcotte EM (2003) Expression deconvolution: a reinterpretation of dna microarray data reveals dynamic changes in cell populations. Proceedings of the National Academy of Sciences, 100, 10370–10375.

Maier MJ (2014) Dirichletreg: Dirichlet regression for compositional data in r. Research Report Series / Department of Statistics and Mathematics, 125.

Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, Williams SR et al. (2020) Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. bioRxiv.

Newman AM, Liu CL, green MR, gentles AJ, Feng W, Xu Y. et al. (2015) Robust enumeration of cell subsets from tissue expression profiles. Nature Methods, 12, 453–457. [PubMed: 25822800]

Newman AM, Steen CB, Liu CL, gentles AJ, Chaudhuri AA, Scherer F. et al. (2019) Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnology, 37, 773–782.

Parikshak NN, Swarup V., Belgard T.g., Irimia M., Ramaswami G., Gandal MJ, et al. . (2016) genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. Nature, 540, 423–427. [PubMed: 27919067]

Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM et al. (2010) Cell type-specific gene expression differences in complex tissues. Nature Methods, 7, 287–289. [PubMed: 20208531]

Sun W, Liu Y, Crowley JJ, Chen T-H, Zhou H, Chu H. et al. (2015) Isodot detects differential rna-isoform expression/usage with respect to a categorical or continuous covariate with high sensitivity and specificity. Journal of the American Statistical Association, 110, 975–986. [PubMed: 26617424]

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C. et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature, 46, 470–476.

Wang N, gong T, Clarke R, Chen L, Shih I-M, Zhang Z. et al. (2014) Undo: a bioconductor r package for unsupervised deconvolution of mixed gene expressions in tumor samples. Bioinformatics, 31, 137–139. [PubMed: 25212756]

Zhong Y, Wan Y-W, Pang K, Chow LM and Liu Z. (2013) Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinformatics, 14, 89. [PubMed: 23497278]

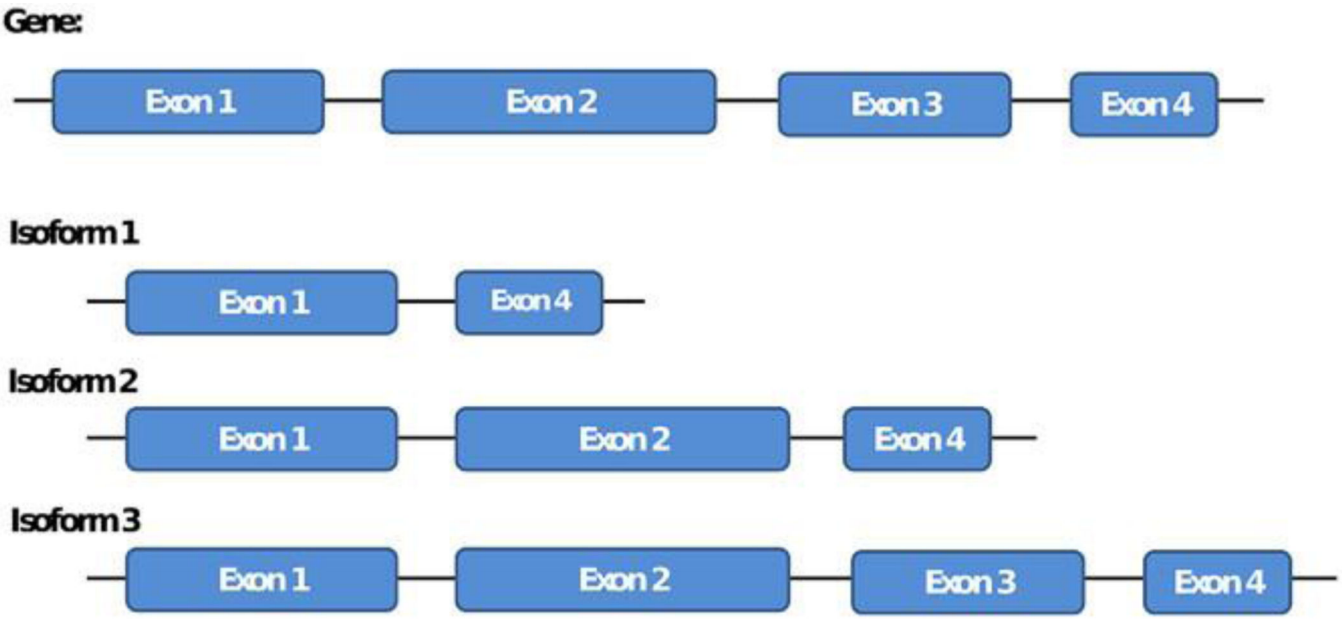

FIGURE 1.
Hypothetical gene and isoform construction model

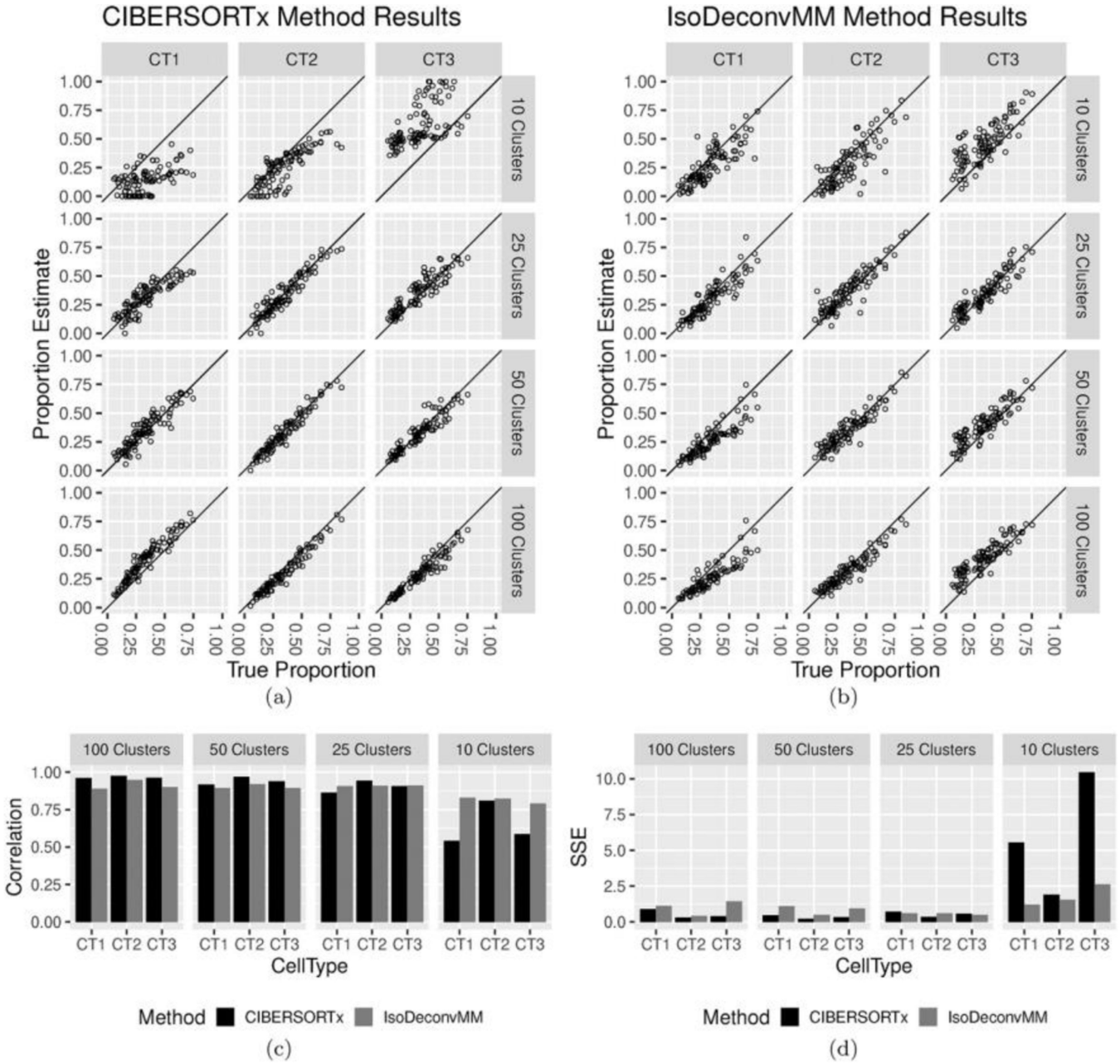

FIGURE 2.

Blueprint mixture proportion estimate results calculated using the CIBERSORTx and IsoDeconvMM methods. Results separated by cell types and number of transcript clusters used in the analysis. (a) Proportion estimates vs true proportions for CIBERSORTx method (used DE clusters only). (b) Proportion estimates vs true proportions for IsoDeconvMM method (used DU clusters only). (c) Correlation and (d) sum-of-square (SSE) results compared across methods

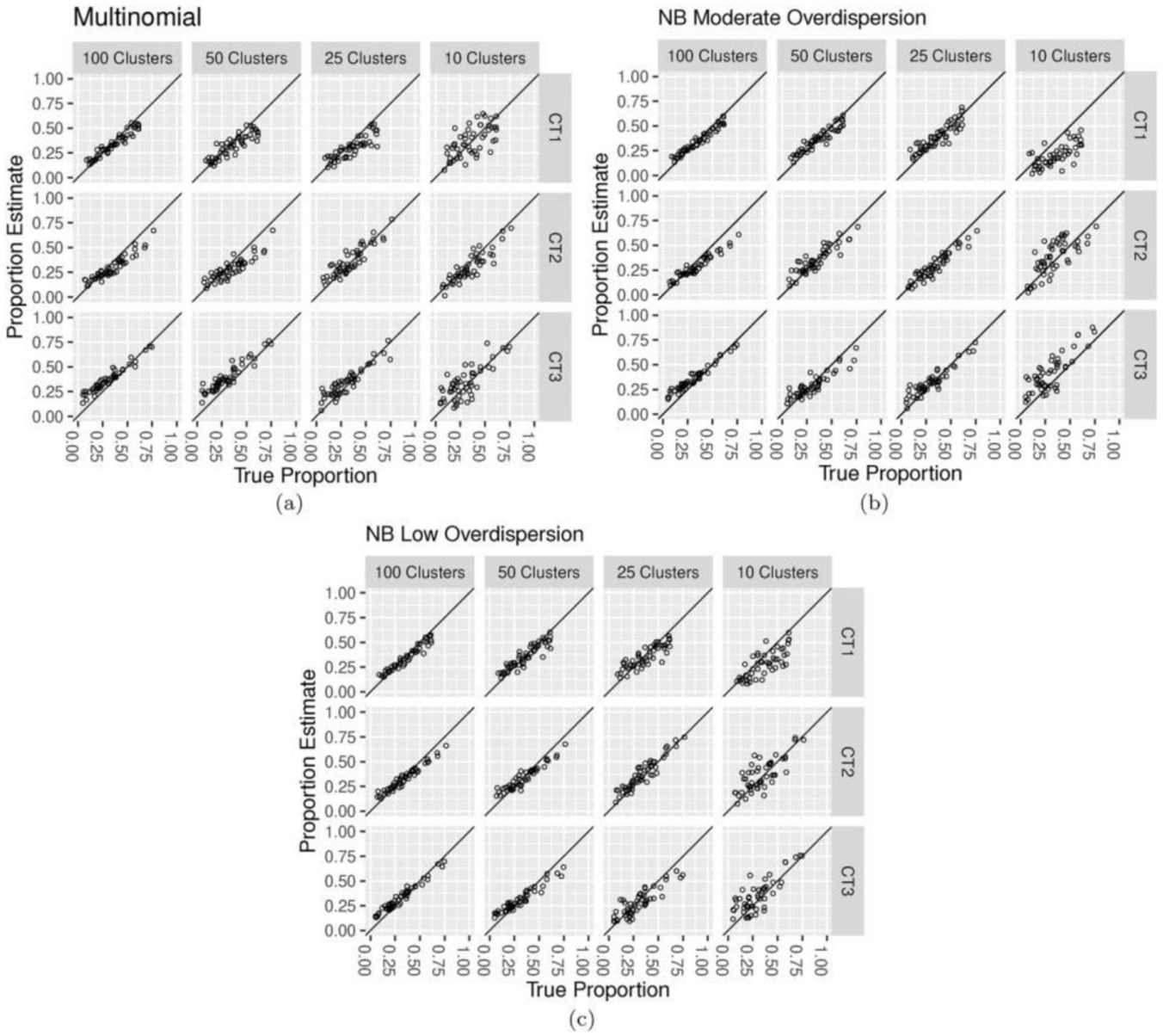

FIGURE 3.

IsoDeconvMM proportion estimates for different underlying data models: (a) Dirichlet-multinomial, (b) Dirichlet-negative binomial with moderate overdispersion, and (c) Dirichlet-negative binomial with low overdispersion. Results separated by cell types (rows) and number of transcript clusters used in the analysis (columns)

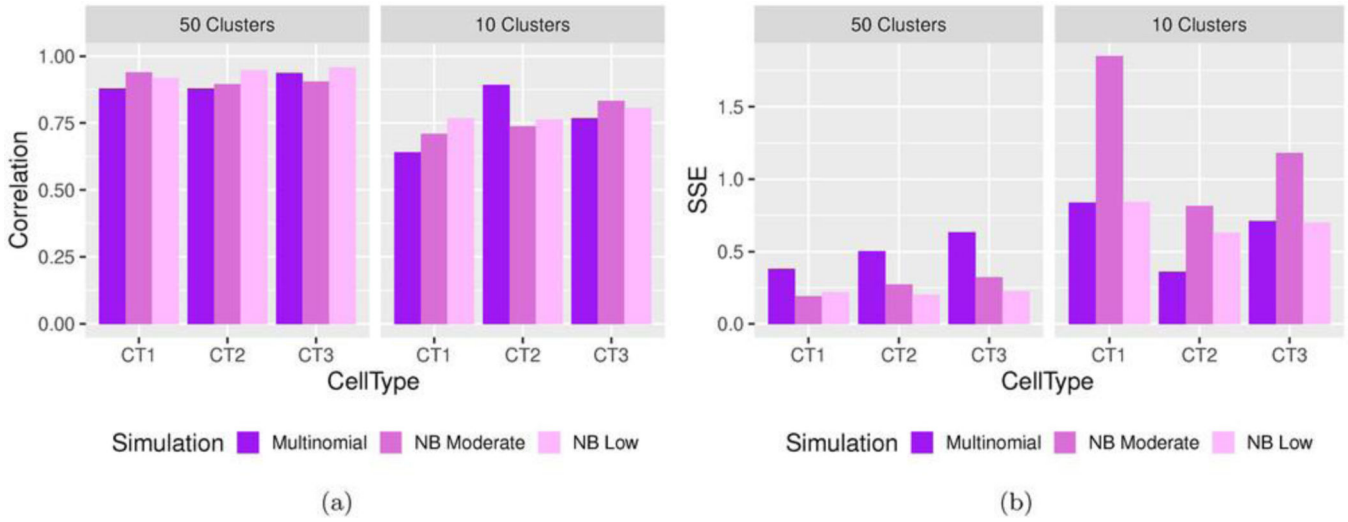

FIGURE 4.

Correlation and sum-of-square error (SSE) results comparing the IsoDeconvMM proportion estimates vs the true proportions for simulations assuming different underlying data models: Dirichlet-multinomial, Dirichlet-negative binomial with moderate overdispersion, and Dirichlet-negative binomial with low overdispersion. Results separated by cell types and number of transcript clusters used in the analysis. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

TABLE 1

The exon sets available for each of the three isoforms from the hypothetical gene in Figure 1. Value of 1 indicates that a paired-end read could theoretically map to that exon set given that the read comes from the isoform specified; value of 0 otherwise

Exon set	Isoform 1	Isoform 2	Isoform 3
$\{E_1\}$	1	1	1
$\{E_2\}$	0	1	1
$\{E_3\}$	0	0	1
$\{E_4\}$	1	1	1
$\{E_1, E_2\}$	0	1	1
$\{E_1, E_3\}$	0	0	1
$\{E_1, E_4\}$	1	1	1
$\{E_2, E_3\}$	0	0	1
$\{E_2, E_4\}$	0	1	1
$\{E_3, E_4\}$	0	0	1
$\{E_1, E_2, E_3\}$	0	0	1
$\{E_1, E_2, E_4\}$	0	1	1
$\{E_1, E_3, E_4\}$	0	0	1
$\{E_2, E_3, E_4\}$	0	0	1
$\{E_1, E_2, E_3, E_4\}$	0	0	1