The Annals of Applied Statistics
2021, Vol. 15, No. 1, 270–286
https://doi.org/10.1214/20-AOAS1395

THE ROLE OF SCALE IN THE ESTIMATION OF CELL-TYPE PROPORTIONS

BY GREGORY J. HUNT[1] AND JOHANN A. GAGNON-BARTSCH[2]

[1]*Department of Mathematics, William & Mary, ghunt@wm.edu*
[2]*Department of Statistics, University of Michigan, johanngb@umich.edu*

Complex tissues are composed of a large number of different types of cells, each involved in a multitude of biological processes. Consequently, an important component to understanding such processes is understanding the cell-type composition of the tissues. Estimating cell-type composition using high-throughput gene expression data is known as cell-type deconvolution. In this paper we first summarize the extensive deconvolution literature by identifying a common regression-like approach to deconvolution. We call this approach the unified deconvolution-as-regression (UDAR) framework. While methods that fall under this framework all use a similar model, they fit using data on different scales. Two popular scales for gene expression data are logarithmic and linear. Unfortunately, each of these scales has problems in the UDAR framework. Using log-scale gene expressions proposes a biologically implausible model and using linear-scale gene expressions will lead to statistically inefficient estimators. To explore ways to address these issues, in this paper we consider how deconvolution methods may use an adjusted model that is a hybrid of the two scales. In analysis on simulations as well as a collection of eleven real benchmark datasets, we find a prototypical hybrid-scale adjustment to the UDAR framework improves statistical efficiency and robustness. More broadly, we believe these hybrid-scale modeling principles may be incorporated into many existing deconvolution methods.

1. Introduction. The tissues of multicellular organisms are typically comprised of a combination of many types of cells. As each cell type has its own set of functions and behaviors, the composition and interaction of different cell types is integral to the function and behavior of the tissues. Thus, studying cell-type composition has long been of broad biological interest. Examples of the importance of cell-type composition abound from the biological literature. In the study of infectious diseases, the composition of white blood cells is important, as it is indicative of many types of dysfunctions (George and Panos (2007)). For example, the number of T-cells among human peripheral blood mononuclear cells (PBMCs) spikes after a Lyme infection (Bouquet et al. (2016)). In neuroscience the composition of brain cells has long been a subject of study. For example, studying the relative composition of microglia in human brains is of interest for those studying developmental dynamics (Ayana, Singh and Pati (2018)). Similarly, understanding changes in the number of neuron and glial cells has been the subject of extensive study with regards to Alzheimer's disease (Mohammadi et al. (2015)).

For this reason, methods to estimate cell-type proportions from high-throughput genomics data have been extensively studied over the past two decades (for comprehensive literature reviews, see Gaujoux (2013) or Mohammadi et al. (2015)). Estimating cell-type proportions is known as *cell-type deconvolution*. Given gene expression data from samples comprised of a mixture of cell types, deconvolution methods estimate the proportions of the constituent cell types. These cell-type proportions may be of interest in their own right, for example, to track

Received March 2020; revised September 2020.
Key words and phrases. Deconvolution, proportions, high-throughput sequencing.

the changes in cell-type composition over time (Newman et al. (2015)). In other cases the estimated cell-type proportions are used as a means of deconfounding differential expression analysis (Capurro et al. (2015)). In this case the cell-type proportions can help explain observed gene-expression differences across samples. By including the estimated proportions in a model, one can separate differences coming from within-cell-type changes in gene expression and those differences coming purely from cell-type-compositional differences among samples (Hagenauer et al. (2016)).

In this paper we present a critique of existing cell-type deconvolution methods and then explore ways to adjust the traditional approach to address the issues we raise. First, in Section 2 we characterize existing deconvolution literature, proposing a new unified deconvolution framework called the unified deconvolution-as-regression (UDAR) framework. The UDAR framework summarizes much of the existing deconvolution literature, including many popular deconvolution methods. It demonstrates that these methods employ a common unified model of the data and, mainly, differ in how their parameter estimates are fit. One important fitting consideration is data scale. Broadly, methods either fit using linear-transformed or log-transformed gene expression data. Unfortunately, each of these scales has problems. We will show that using log-scale gene expressions proposes a biologically implausible model and that using linear-scale gene expressions will lead to statistically inefficient estimators. Using the UDAR framework as a point of comparison, in Section 3 we consider how this framework may be modified to take advantage of the beneficial aspects of fitting on a hybrid of the two scales. Subsequently, in Section 4 we explore the differences between existing approaches and an example prototypical hybrid-scale model. We do this across a wide range of simulated data as well as 11 real benchmark datasets. Here, we see that models using an adjusted hybrid scale have improved statistical efficiency and robustness.

2. A unified framework for existing deconvolution models. Let $Y \in \mathbb{R}^N$ be the measurements of N gene expressions in a mixture sample of K types of cells and $R \in \mathbb{R}^{N\times K}$ be reference expressions of these N genes across the K constituent cell types. Furthermore, let $p = (p_1, \ldots, p_K)$ be the proportions of the K cell types in the mixture sample. Implicit in being proportions is that p must satisfy the sum-to-one (STO) constraint: $\sum_{k=1}^K p_k = 1$ and the nonnegativity (NN) constraint: $p_k \geq 0$ for $k = 1, \ldots, K$. That is, $p \in \Delta_{K-1}$, the $(K-1)$ probability simplex $\Delta_{K-1} = \{x \in \mathbb{R}^K : x_k \geq 0 \text{ and } \sum_{k=1}^K x_k = 1\}$.

The deconvolution problem is that p is unknown, and we want to estimate it. In this section we introduce a new unified model for cell-type deconvolution called the unified deconvolution-as-regression (UDAR) framework. The UDAR framework posits that Y, R and p are related through the linear model

$$Y = Rp + \varepsilon \tag{1}$$

for a random error ε. Estimating p under this model is equivalent to solving a constrained regression of Y on R where the coefficients p must live in Δ_{K-1}. Hence, using this model is treating deconvolution as regression. Note that we only consider problems where Y and R are known, and we are interested in estimating p. We do not consider the related problem where R is also unknown. For a discussion of this problem, see Gaujoux (2013), Mohammadi et al. (2015) or Wang et al. (2016).

What differs among existing deconvolution methods is the approach by which p is estimated. There are common themes among how estimates of $\hat{p}$ are fit. Typically, methods specify: (1) a loss function $L : \mathbb{R}^K \to \mathbb{R}_+$ that determines model fit $L(p)$ for putative proportions p, (2) an optimization space $\Pi \subseteq \mathbb{R}^K$ for p and (3) a post hoc adjustment function $\varphi : \Pi \to \Delta_{K-1}$ mapping from the optimization space Π to the desired simplex Δ_{K-1}. They then estimate p by minimizing L over Π and applying φ. This approach is described in Algorithm 1. The idea behind this approach is that, while, ideally, $\hat{p}$ is the minimizer of L over

Algorithm 1 UDAR Fitting

Step 1: Minimize L over Π to get p^*:

$$p^* = \arg\min_{p \in \Pi} L(p)$$

Step 2: Apply φ to p^* to get $\hat{p}$:

$$\hat{p} = \varphi(p^*).$$

Δ_{K-1}, solving such a constrained minimization problem is difficult. Thus, UDAR methods solve an easier relaxation of this problem, minimizing L over $\Pi \supseteq \Delta_{K-1}$ and then making post hoc adjustments to p^* to produce a final estimate $\hat{p} \in \Delta_{K-1}$.

A large number of existing deconvolution methods fit into this framework under appropriate choices of L, Π and φ. The most common choice of loss is the squared-error loss (Abbas et al. (2009), Gong and Szustakowski (2013), Gong et al. (2011), Lu, Nakorchevskiy and Marcotte (2003), Qiao et al. (2012), Racle et al. (2017), Wang, Master and Chodosh (2006)). Other loss functions used include an elastic net penalized loss (Altboum et al. (2014)), a support-vector regression approach, which is equivalent to using an ϵ-insensitive loss (Newman et al. (2015)), a Bayesian-likelihood approach based on latent Dirichlet allocation that is equivalent to letting L be a likelihood-based loss (Blei, Ng and Jordan (2003), Qiao et al. (2012)), a negative binomial likelihood (Du, Carey and Weiss (2019)), KL-divergence based method (Li (2019)), robust regression approaches (Finotello et al. (2019)) and a gene-set-based component merging approach (Wang et al. (2018)). The optimization space Π is typically one of three spaces: (1) Δ_{K-1} (Du, Carey and Weiss (2019), Finotello et al. (2019), Gong and Szustakowski (2013), Gong et al. (2011)), (2) $\mathbb{R}^K_+$, the positive orthant of $\mathbb{R}^K$ (Li (2019), Qiao et al. (2012), Racle et al. (2017)) or (3) $\mathbb{R}^K$ (Abbas et al. (2009), Lu, Nakorchevskiy and Marcotte (2003), Newman et al. (2015), Wang, Master and Chodosh (2006)). In the first case, where $\Pi = \Delta_{K-1}$, no post hoc adjustments are necessary and so φ is the identity function. In the second case where $\Pi = \mathbb{R}^K_+$, since p^* already satisfies the NN constraint, φ re-normalizes p^* to enforce the STO constraint and hence $\hat{p}_k = \varphi(p^*)_k = p^*_k / \sum_{t=1}^K p^*_t$ (Qiao et al. (2012), Racle et al. (2017)). Finally, in the first case of unconstrained optimization where $\Pi = \mathbb{R}^K$, φ zeros out negative coefficients and then renormalizes so that $\hat{p}_k = \varphi(p^*)_k = (p^*_k)_+ / \sum_{t=1}^K (p^*_t)_+$ where $(\cdot)_+ = \max(\cdot, 0)$ is the positive part (Abbas et al. (2009), Lu, Nakorchevskiy and Marcotte (2003), Newman et al. (2015), Wang, Master and Chodosh (2006)). We call this latter post hoc adjustment the "zero-then-renormalize" adjustment.

2.1. *Scale considerations for deconvolution.* An important question for the UDAR framework is the appropriateness of this model for the deconvolution problem. One important modeling consideration is data scale. Typically, gene expressions are either linearly transformed, for example, TPM (Conesa et al. (2016)), or logarithmically transformed, for example, RMA, (Irizarry et al. (2003)). In the former case we say the data is on the linear scale and in the latter we say the data is on the log scale. Some deconvolution methods assume linear-scale expressions, like in Newman et al. (2015), some methods assume log-scale expressions, as in Qiao et al. (2012); most make no explicit assumptions about data scale at all. In the following sections we will consider the appropriate data scale for the UDAR model. This will primarily concern the two major components of the model: (1) the linear mean structure Rp and (2) the additive error structure ε.

2.1.1. *Mean modeling.* Assume we have a mixture sample comprised of cell types $k = 1, \ldots, K$ in proportions $p_1, \ldots, p_K$. First, notice that if η_n is the amount of mRNA in our mixture sample coming from gene n and η_{nk} is the amount of that mRNA in the sample coming from type k cells, then

$$\eta_n = \sum_{k=1}^{K} \eta_{nk}. \tag{2}$$

Now, assume we also have some reference sample of type k cells. Let the amount of mRNA from gene n in the reference sample be η^*_{nk}. Since the mixture sample is comprised of a proportion p_k of type k cells and the reference sample is 100% type k cells, then we expect that

$$\eta_{nk} \approx p_k \eta^*_{nk}, \tag{3}$$

assuming the same abundance of cells between the reference samples and the mixture. Essentially, this assumes that type k cells in the mixture behave as if they were a random sample of the type k reference cells. We assume that this relationship is only approximate because the type k reference cells may not exactly mimic the type k mixture cells. For example, the microenvironment of the cells in the mixture may modify gene expression.

Combining equations (2) and (3), we get that

$$\eta_n = \sum_{k=1}^{K} \eta_{nk} \approx \sum_{k=1}^{K} p_k \eta^*_{nk}. \tag{4}$$

Now, assume that the *linear scale* measured gene expressions are proportional to the amount of mRNA so that $Y_n \approx \gamma \alpha_n \eta_n$ and $R_{nk} \approx \alpha_n \eta^*_{nk}$ for constants γ and $\{\alpha_n\}$. The proportionality constants α_n capture gene-specific effects like probe affinity (for microarray data) or length biases (for RNA-seq). The multiplier γ captures global differences between the mixture and references. This includes effects like sequencing depth or amount of mRNA. Again, we assume approximate equality because the measurement process may introduce random errors. Combining with equation (4) we now get that

$$Y_n \approx \gamma \alpha_n \eta_n \approx \gamma \alpha_n \sum_{k=1}^{K} p_k \eta^*_{nk} = \gamma \sum_{k=1}^{K} p_k \alpha_n \eta^*_{nk} \approx \gamma \sum_{k=1}^{K} p_k R_{nk}.$$

As is customary, assume that Y and R have been normalized to account for global expression differences, for example, by TPM (Conesa et al. (2016)), so that $\gamma = 1$. Then, the above equation shows that the linear model $Y \approx Rp$, proposed by UDAR, is correctly specified for linear-scale gene expression measurements since $Y_n \approx \sum_{k=1}^{K} p_k R_{nk}$.

However, even if $\gamma = 1$, the linear mean structure is misspecified for log-scale gene expressions as

$$\log(Y_n) \approx \log(\alpha_n \eta_n) \approx \log\left(\sum_{k=1}^{K} \alpha_n p_k \eta^*_{nk}\right) \not\approx \sum_{k=1}^{K} p_k \log(\alpha_n \eta^*_{nk}) \approx \sum_{k=1}^{K} p_k \log(R_{nk})$$

since we can't interchange a sum and a log. Thus, $\log(Y_n) \not\approx \sum_{k=1}^{K} p_k \log(R_{nk})$, and so a linear mean structure, as proposed by UDAR, does not make sense on the log scale. For a toy example of this principle, see Figure 1.

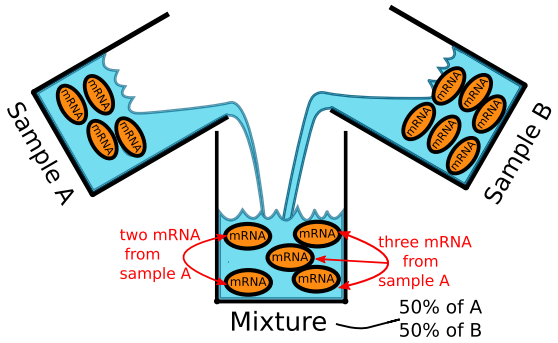

FIG. 1. *The mixture sample is* 50% *of A and* 50% *of B. The orange ovals represent mRNA from a specific gene. Since the reference of type A typically has four mRNA, we expect* $4 \times .5 = 2$ *mRNA in the mixture to come from ref. A. Similarly, since ref. B typically has six mRNA, we expect* $6 \times .5 = 3$ *mRNA in the mixture to have come from ref. B. In total, we get* $5 = 4 \times .5 + 6 \times .5$ *mRNA in the mixture. Thus, the amount of mRNA in the mixture is a linear mixture of the amount of mRNA. This does not work if we logarithmically transform the counts. In that case we would expect, on the log scale, to get* $\log(4) \times .5 + \log(6) \times .5 \approx 1.6$ *mRNA. Exponentiating back to the linear scale, this is* ≈ 4.9*, thus under-counting the true amount of mRNA.*

2.1.2. *Error modeling.* In contrast to the mean structure, error assumptions are most reasonable for log-scale expressions. While most methods simply note that $Y \approx Rp$ and do not explicitly include an error term ε in their models, their loss functions are optimal for typical regression-like error assumptions about ε. For example, deconvolution methods minimizing the squared-error loss are are optimal when $\varepsilon_n \overset{\text{iid}}{\sim} N(0, \sigma^2)$ with some constant error variance $\sigma^2 > 0$. Such regression assumptions are most appropriate on the log scale. Indeed, it has been widely noted that errors are well modeled as normal with approximately constant variance across expression levels for log-scale gene expression data (Qiao et al. (2012)). Conversely, error for linear-scale expression data are right-skewed, and the variance tends to increase with increasing mean expression (Hardin and Wilson (2009), Qiao et al. (2012), Tu, Stolovitzky and Klein (2002), Weng et al. (2006), Zwiener, Frisch and Binder (2014)).

3. A hybrid model for deconvolution. The previous two sections present a problem for many existing deconvolution methods. If they follow the UDAR model on the log scale, they will have a misspecified mean. Conversely, if they propose the UDAR model with linear-scale expressions, the error assumptions are unrealistic. While nearly all existing methods model deconvolution on one of these two scales, two exceptions exist: Hunt et al. (2019) and Wilson et al. (2020). These methods attempt to deal with the scale considerations by proposing linear mixing with log-scale errors. In this section we distill the central ideas of all of these approaches and explore a straightforward way to augment the UDAR model in a way that avoids the problems of scale while not making restrictive assumptions. In Section 4 we will use simulations and real data to examine the performance of this prototypical hybrid approach alongside other log-scale, linear-scale, and the two existing hybrid-scale approaches.

As a prototypical example of a hybrid-scale method, we will work with the following model:

$$\log(Y_n) = \theta + \log\left(\sum_{k=1}^{K} R_{nk} p_k\right) + \varepsilon_n, \tag{5}$$

where $\varepsilon_n \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Again, Y_n and R_{nk} denote the linear-scale gene expression. This model proposes additive Gaussian error *after* a log transformation and thus uses an appropriate scale for errors. Furthermore, the mean structure in equation (5) implies $\log(Y) \approx \theta + \log(Rp)$, or,

equivalently, $Y \approx e^{\theta} Rp$. Thus, it proposes a plausible linear-mixing structure on the *linear scale*, as discussed in Section 2.1.1. The term e^{θ} plays the same role as the term γ mentioned in the previous section and accounts for systematic differences between the mixture and reference expressions, such as differences in sequencing depth or total RNA content.

To estimate p under this model, we let $\hat{p}$ be the MLE so that

$$(\hat{p}, \hat{\sigma}^2, \hat{\theta}) = \arg\min_{p \in \Delta_{K-1}, \sigma^2 \in \mathbb{R}_+, \theta \in \mathbb{R}} \ell(p, \sigma^2, \theta) \tag{6}$$

and ℓ is the joint log-likelihood function of p, σ^2, and θ. In the next subsection we present a novel way to find the MLE for this model by minimizing a variance-based loss. Thus, the optimization problem in equation (6) can be solved using an approach analogous to the UDAR fitting procedure. However, more generally, one could use the model in equation (5) and estimate p by minimizing other loss functions. For example, one might consider an L2-penalized loss if there are high correlations among cell types (Altboum et al. (2014)). Alternatively, one might use an L1 penalized loss if one believes many of the cell types to not be truly present in the sample. Other losses, like the SVR-based approaches used in Newman et al. (2015) and Fernández et al. (2019), have similar advantages by promoting sparsity in the solutions. This can prove useful when there are many potential cell types to include in the analysis but reason to believe that only a few of the cell types are actually present. By promoting a sparse solution and forcing certain estimates to be exactly zero, such losses can help determine the true number of cell types present and more accurately estimate the constituent proportions. Broadly, estimation under a hybrid-scale framework, like that in equation (5), is general and extensible in many of the same ways as the UDAR framework while allowing proper modeling of the mean and error structures. Thus, we believe that many of these other innovative approaches could compatibly be incorporated into this hybrid-scale model in future work.

3.1. *A method for finding the MLE.* For this hybrid model there is a UDAR-like procedure analogous to Algorithm 1 to find the MLE $\hat{p}$. Define $\lambda_n(p) = \log(Y_n) - \log(\sum_{k=1}^{K} R_{nk} p_k)$ as the intercept-free residual (i.e., the residual ignoring θ), and let $S^2(p)$ be the sample variance of the $\lambda_n(p)$, $S^2(p) = N^{-1} \sum_{n=1}^{N} (\lambda_n(p) - \bar{\lambda}(p))^2$ where $\bar{\lambda}(p) = N^{-1} \sum_{n=1}^{N} \lambda_n(p)$. It can be shown (see Supplementary Material, Section 1) that $\hat{p}$ is the minimizer of S^2 so that

$$\hat{p} = \arg\min_{p \in \Delta_{K-1}} S^2(p).$$

Furthermore, since S^2 is invariant under scaling so that $S^2(cp) = S^2(p)$ for any $c \in \mathbb{R}_+$, we do not need to optimize over Δ_{K-1} directly. Instead, we can solve this optimization problem over any positive set containing Δ_{K-1} and simply renormalize. Let p^* be any minimum of $S^2(p)$ over $p \in \Pi$ where Π is any set satisfying $\Delta_{K-1} \subseteq \Pi \subseteq \mathbb{R}_+^K$ so that $p^* = \arg\min_{p \in \Pi} S^2(p)$. (Notice that this minimum is not unique since if p^* minimizes S^2 then so does cp^*.) Then, if $T^* = \sum_{k=1}^{K} p_k^*$ is the sum of the elements of p^*, the MLE for p is $\hat{p} = p^*/T^*$ since $\hat{p} \in \Delta_{K-1}$. This is summarized in Algorithm 2, a straight-forward procedure to estimate $\hat{p}$.

This procedure allows us to find the MLE without trying to minimize ℓ over p, σ^2 and θ simultaneously. Furthermore, like the UDAR model, this procedure also allows us to optimize p over a relaxation $\Pi = [0, 1]^K \supseteq \Delta_{K-1}$ instead of having to directly search over Δ_{K-1}. Also, like Algorithm 1, we renormalize p^* to form a $\hat{p}$ that is in Δ_{K-1}. However, while for the UDAR framework the post hoc adjustments were a heuristic to enforce constraints on $\hat{p}$, our renormalization is not heuristic. The two steps in Algorithm 2 precisely recover the MLE of the hybrid model without solving a difficult optimization problem over Δ_{K-1}. While we could have optimized $S^2(p)$ over any space Π where $\Delta_{K-1} \subseteq \Pi \subseteq \mathbb{R}_+^K$, letting $\Pi = [0, 1]^K$ greatly simplifies the optimization problem and allows us to use standard global optimization routines with box constraints to find $\hat{p}$.

Algorithm 2 Hybrid Fitting Procedure

Step 1: Minimize S^2 over the parameter space $\Pi = [0, 1]^K \supseteq \Delta_{K-1}$ to get p^*:

$$p^* = \arg\min_{p \in \Pi} S^2(p)$$

Step 2: Form $\hat{p}$ by modifying p^* to ensure it satisfies the sum-to-one (STO) constraint by defining

$$\hat{p} = p^* / T^*$$

where $T^* = \sum_{t=1}^{K} p_t^*$.

3.2. *References, marker genes and weights.* Reference data is typically obtained from online gene expression repositories, like GEO (Edgar (2002)), or from specific profiles complied for cell-type deconvolution. Such reference data is used in two major ways (Gaujoux (2013)). First, the reference data is used to create the reference matrix R so that R_{nk} is the typical expression of gene n in a sample purely of cells of type k. If there exists more than one reference sample for a particular cell type, one typically averages the profiles. If one has ν_k reference profiles of cell type k, then R_{nk} is typically average expression across the profiles so that $R_{nk} = (\nu_k)^{-1} \sum_{r=1}^{\nu_k} R_{nkr}$ where R_{nkr} is the gene expression of gene n in the rth reference of cell type k.

In addition to using reference data to form the reference matrix R, this reference data is often used to find marker genes. Marker genes are genes that are particularly highly expressed in one cell type but not the others. Typically, marker genes are identified by comparing gene expression across cell types in the reference data using, for example, a t-test. Once identified, deconvolution methods typically fit using only the subset of marker genes. Let $\mathcal{M} \subseteq \{1, \ldots, N\}$ be the set of marker genes. Then, the use of marker genes can be viewed as variable selection where we only fit using those $n \in \mathcal{M}$. Alternatively, we can view the marker genes as a weighting of the loss function. Under the UDAR model, fitting using $\mathcal{M}$ is equivalent to using a weighted loss function with weights $w_n = \mathbb{1}(n \in \mathcal{M})$.

In the case where we have more than one reference of each cell type, we may also do so in a weighted fashion, weighting inversely with estimated variance. This enables incorporation of variance information from the reference data if it is available. Hybrid-scale models can also encompass marker genes as variable selection or a weighted loss. For example, optimizing the loss over only those $n \in \mathcal{M}$.

4. Results.

4.1. *Comparison of methods on simulated data.* To explore the properties of hybrid models, as compared to the UDAR model, we first consider simulated mixtures data. We simulate mixtures using reference RNA-seq profiles of brain, liver and muscle cells from Parsons et al. (2015). Let $R \in \mathbb{R}^{N \times K}$ as the reference profile matrix of the $N = 23{,}459$ genes profiled in the $K = 3$ reference samples so that R is comprised of linear scale (untransformed) read counts. We generate mixture proportions p uniformly from Δ_{K-1} and form a simulated mixture profile $Y \in \mathbb{R}^N$ so that

$$\log(Y_n) \overset{\text{iid}}{\sim} N\big(\log\big((Rp)_n\big), \tau D^2\big), \tag{7}$$

where $D^2 \approx 1.2$ is the median within-gene sample variance in the reference data and τ is a variance multiplier parameter we are free to choose.

In this section we consider five approaches to deconvolution. We use the hybrid model following Algorithm 2, two example OLS-based UDAR approaches (one on the linear-scale

TABLE 1
Five Methods compared in this section. The Hybrid scale, two OLS UDAR approaches, two existing UDAR approaches from the literature

Method	Scale	Loss	Π
Hybrid	Hybrid	Variance	Δ_{K-1}
Regression	Linear	L2	$\mathbb{R}^K$
Log. Regression	Log	L2	$\mathbb{R}^K$
cibersort	Linear	ε-insensitive	$\mathbb{R}^K$
deconvSeq	Log	Neg. Bin. Likelihood	Δ_{K-1}

and log-scale, respectively) and two recent methods from the literature: a linear-scale support-vector regression approach called cibersort (Newman et al. (2015)) and a log-scale negative binomial regression called deconvSeq (Du, Carey and Weiss (2019)). All but the hybrid approach follow the UDAR framework to solve the regression problem. The OLS approaches minimize a squared-error loss L, optimizing over $\Pi = \mathbb{R}^K$, and apply the simple zero-then-renormalize post hoc adjustments. We call the linear-scale version a "Regression" approach because it is equivalent to letting $\hat{p}_k = (p_k^*)_+ / \sum_{t=1}^K (p_t^*)_+$ where p^* are the coefficients obtained from regressing Y on R. We call the log-scale version approach "Log Regression" because it is equivalent to the regression approach, but where the p^* are the coefficients obtained from regressing $\log(Y)$ on $\log(R)$, cibersort falls under the UDAR framework using an ε-insensitive loss (i.e., using support-vector regression) and optimizing over $\Pi = \mathbb{R}^K$. It then applies the zero-then-renormalize post hoc adjustment. deconvSeq minimizes a negative binomial likelihood loss over $\Pi = \Delta_{K-1}$. In application of deconvSeq to continuous data, linear-scale expressions were necessarily rounded. For all methods we subset Y and R to a set of marker genes chosen by an ANOVA on the reference data. We will let M denote the number of marker genes used for each cell type. The exact same set of marker genes are used to fit the methods. These methods are summarized in Table 1.

In Figure 2 we plot scatter plots of the estimates against the truth for each of the methods. There are four subplots for three different simulation settings. In each we set $\tau = \frac{1}{2}$ (low noise) and then vary the number of markers over $M = 10$, 100 and 1000. For each setting and method we estimate the proportions for 50 simulated samples. From these plots we can see that the hybrid-scale approach generally out-performs the other approaches. The log regression and deconvSeq do comparatively poorly because they have a misspecified mean, and thus an obvious bias manifested in the S-shaped relationship between the truth and the estimates. The other three approaches do not exhibit this bias. On average, their estimates generally track the true mixing proportions. Nonetheless, linear-scale regression and cibersort both perform worse than the hybrid approach because they have higher variance. Thus, the estimates for the hybrid approach are typically closer to the truth than for the other two linear-scale methods (regression and cibersort). The linear-scale regression used by regression and cibersort produce statistical inefficiencies evidenced by the higher variance. In the Supplementary Material Figure 1 we display plots for the four methods over a larger range of simulation settings and more methods including robust regression, weighted regression, constrained regression, a negative-binomial GLM and other methods from the literature. We also include similar plots in the Supplementary Material Figure 2 using as reference 22 white blood cell types from (Newman et al. (2015)). This demonstrates a scenario with many closely-related cell types. These plots show largely the same story observed in Figure 2.

To explore the role of the Gaussianity assumption on performance, in Figure 3 we construct similar scatter plots for data simulated using a negative binomial model for the expressions.

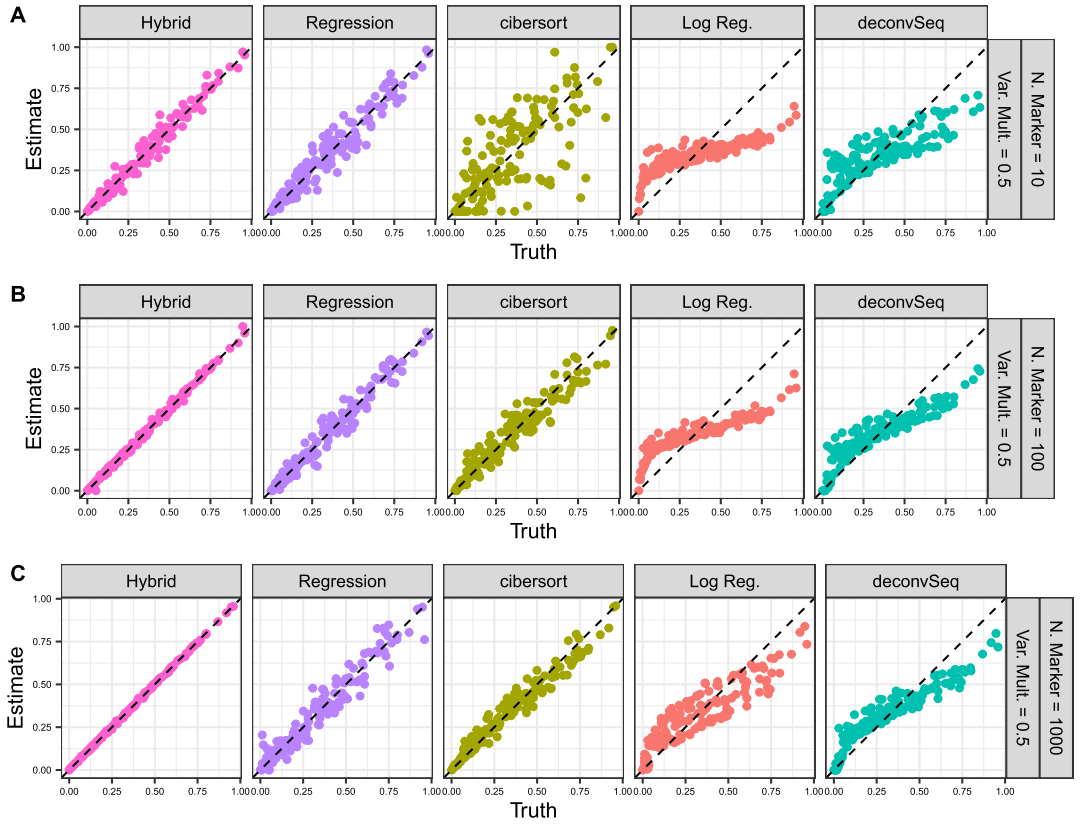

FIG. 2. *Evaluation of methods on simulated mixture data with Gaussian noise for number of marker genes (per cell type) (M). (A) $M = 10$, (B) $M = 100$, (C) $M = 1000$.*

The simulations are similar to those in equation (7); however, we let

$$Y_n \overset{\text{iid}}{\sim} \text{NegBinom}\left(\text{mean} = (Rp)_n, \text{size} = \frac{1}{\psi}\right)$$

so that Y_n has mean $\mu = \mathbb{E}[Y_n] = (Rp)_n$ and variance $\text{Var}(Y_n) = \mu + \mu^2\psi$ where ψ is a parameter we are free to choose. In Figure 3 we consider simulation settings for $\psi = \frac{1}{2}$ (low noise) and vary M over 10, 100, 1000 markers. We see similar behavior for the negative binomial simulations as in the Gaussian case. The hybrid approach outperforms the other approaches, suggesting that the model is relatively insensitive to an exact Gaussian error assumption. The hybrid-scale approach performs better than the other approaches because it uses reasonable scales for both the mean structure and the errors. This leads to both a lower bias and lower variance than the other methods. In the Supplementary Material Figure 3 we display plots of errors for the negative binomial simulations over a wider range of simulation settings. As previously, we also include similar plots in the Supplementary Material Figure 4 using as reference the 22 white blood cell types from (Newman et al. (2015)) to demonstrate efficacy in a setting with many closely-related cell types. These figures tell much the same story.

4.2. *Comparison to other hybrid-scale approaches.* To date, two existing approaches try to model deconvolution as a hybrid of the linear and log-scale expressions. These methods include previous work of the authors developing a method called dtangle (Hunt et al. (2019)) as well as work by Wilson et al. (2020) developing an approach called ICeD-T. Broadly,

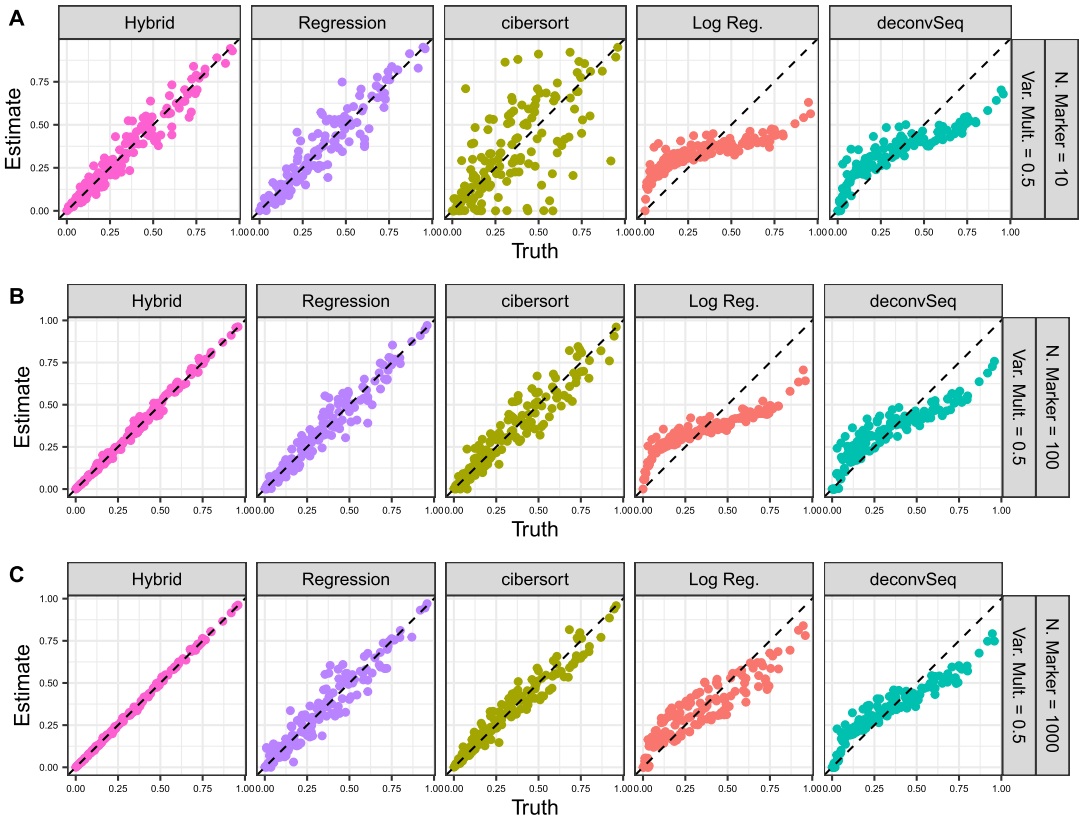

FIG. 3. *Evaluation of methods on simulated mixture data with negative binomial noise for number of marker genes (per cell type) (M): (A) $M = 10$, (B) $M = 100$, (C) $M = 1000$.*

these methods attempt to deal with the scale considerations by proposing linear mixing with log-scale errors. To solve the resulting model for p, each method makes simplifying approximations which place limitations on the applicability of the methods. For dtangle a simplifying assumption is made with regard to marker genes. The working approximation is that the expression of marker genes is exactly zero in all but the cell type they mark. This yields a simple, closed-form analytical estimator of cell-type proportion. This method works well when the assumption is true, that is, as long as not too many genes are included in analysis that do not follow this definition of a marker gene.

The ICeD-T method follows primarily from two modeling choices. First, it approximates the sum of log-normally distributed random variables as log normal. This allows approximate computation of an otherwise nonanalytical likelihood. Second, ICeD-T deals with marker genes by modeling the likelihood as a mixture of two likelihoods: one for genes whose variance is reasonably small and one for genes with abberantly high variance. These two modeling choices ultimately produce a method that incorporates estimates of the error variance for each gene into the parameter updates of an EM algorithm. We will show that this method works well so long as these error variances are low. Note that ICeD-T often fails to converge for many simulations and real datasets. To get ICeD-T to run in as many scenarios as possible, we have made slight modifications to the code to better handle missing values and nonconvergent likelihood optimization. These changes can be summarized in Section 3 of the Supplementary Material.

To compare dtangle and ICeD-T to the hybrid approach introduced in this paper, we generate simulation data to investigate performance on data comprised of highly similar cell types.

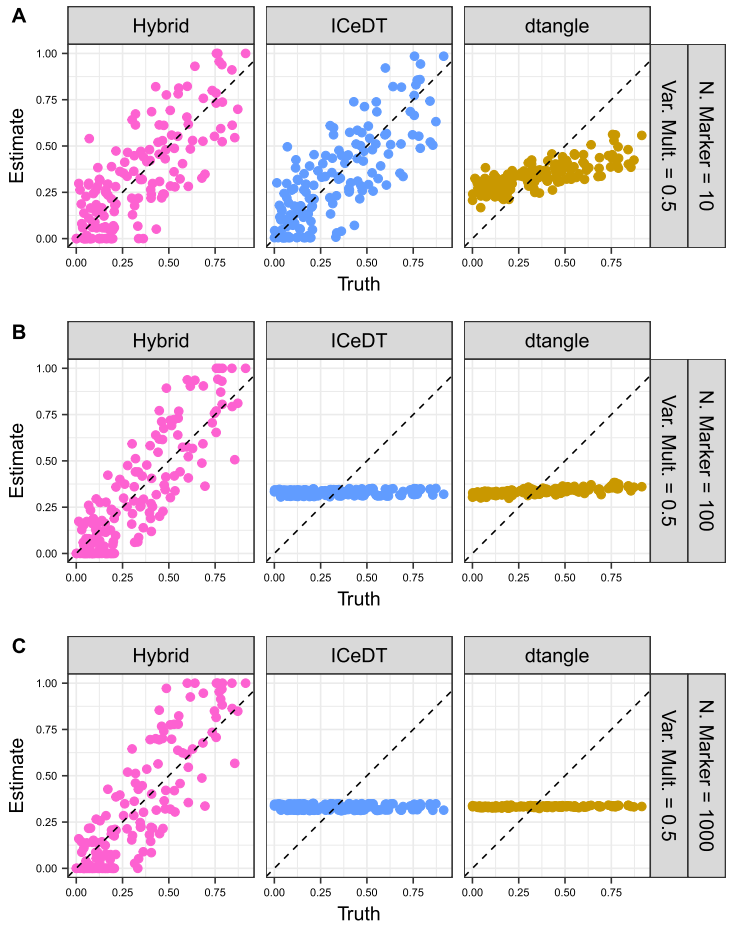

FIG. 4. *Evaluation of hybrid methods on simulated mixture data with Gaussian noise for different using* (A) $M = 10$, (B) $M = 100$, (C) $M = 1000$. *The error variance multiplier is set to* $\tau = .5$.

We generate the data similarly to Figure 2 but we: (1) limit the true number differentially expressed marker genes across the cell types and (2) limit the amount of differential expression of markers across cell types. To begin, we let all of the cell types have an identical expression profile. We then generate marker genes by setting the expression in two of the cell types as the median expression across all genes and then set the expression in the third type as twice this median (on the linear scale). For each cell type we generate 20 marker genes in this way. In total, there are thus 60 marker genes each with a $\log_2$-fold change of one. All other genes in the simulated data are identically expressed. In Figure 4 we explore the performance of the hybrid approach, dtangle and ICeD-T on this data using $M = 10, 100$ and 1000 putative marker genes (per cell type).

This simulation demonstrates a difficult deconvolution problem due to the similarity of the cell types. There are only 60 marker genes in total, and the difference in expression across the markers is only two-fold. We can see from this figure that both dtangle and ICeD-T are very sensitive to the quality of marker genes. As we increase the number of nonmarker genes included in the analysis, both dtangle and ICeD-T converge to estimating all cell types as equally abundant. Since there are three cell types, they predict $\frac{1}{3}$ for each type in each sample. Notice that the hybrid method is relatively insensitive to including genes that are poor quality markers. Its performance does not deteriorate by including nonmarker genes like dtangle or ICeD-T. Importantly, the hybrid method also does not require the marker genes to

be expressed in only one cell type, with zero expression in the other cell types. All that is required is that the marker genes are differentially expressed between the cell types.

4.3. *Supplementary simulations*. In the Supplementary Material, Figures 1 and 3, we plot Gaussian and negative binomial simulations similar to Figures 2 and 3 but over all simulation settings $\tau = \frac{1}{5}, 1, 5$, number of markers $M = 10, 100, 1000$ (per cell type) and include more methods in the analysis. We also make these plots in Figures 2 and 3 of the Supplementary Material using as reference 22 types of white blood cells for comparison on many closely-related cell types. Similarly, in Supplementary Material Figure 5 we plot simulations similar to Figure 4 but over all simulation settings and more methods. Broadly, we see that the hybrid-scale approach performs competitively with existing approaches.

4.4. *Comparison of methods on real data*. To explore deconvolution performance on real data, we use a collection of existing deconvolution benchmark datasets (see Supplementary-Material Table 1). In all, these 11 datasets cover a range of realistic deconvolution settings. Across the datasets there is a range of cell types, number of cell types, organisms (human and rat) and technologies (RNA-seq and microarrays). Some datasets contain reference data created as part of the same sequencing experiment, while other datasets contain third-party references. For most of the datasets, the true mixing proportions are known because the cells were mixed in known proportions before expressions were assayed. However, for three of the datasets the true proportions are the cell-type proportions reported by a physical sorting technique applied after the gene expression assays. Over the past 20 years more than 60 deconvolution methods have been developed (Li (2019)). To broaden the scope of our analysis and in addition to the methods above, we include an additional three recent and popular linear-scale UDAR approaches: deconRNAseq, EPIC and MOMF (Gong and Szustakowski (2013), Li (2019), Racle et al. (2017)) as well as a simplex-constrained regression approach (Constr. Reg.), robust regression approach (RLM), weighted regression approach (Wtd. Reg.) and a negative binomial GLM.

The choice of marker genes is an extremely important component in the application of cell-type deconvolution methods, as accuracy is strongly influenced by the choice of markers. For example, consider Figure 5. In this figure we plot the error for the dataset from Gong et al. (2011) for the four methods. Error is measured as absolute value of the difference between the true proportions and their predictions. We use the exact same marker genes for each method but estimate the cell-type proportions using a range of different numbers of markers per cell type (M). We let M vary following an approximate exponential sequence $M = 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000$, approximately doubling the number of marker genes (per cell type) at each step. We can see from this figure that estimation performance depends heavily on the number of marker genes used. For example, cibersort does poorly for a small number of markers but sees improvement for a large number of markers until it hits a point of diminishing returns and starts to become less accurate for too many markers. Conversely, methods like dtangle, deconRNAseq, EPIC and MOMF (generally) have increased accuracy as the number of marker genes increases.

Importantly, the optimal number of marker genes depends as much on the particular dataset as on the particular method. As an example, consider accuracy as a function of number of marker genes in Figure 6 for the data from Liu et al. (2015). Here, we see that, generally, methods like deconvSeq, dtangle and MOMF have worse error as the number of marker genes increases. Conversely, for methods like cibersort and deconRNAseq we see that the accuracy is increasing as the number of markers increases for this dataset. We display similar plots for the other datasets in the Supplementary Material, Figures 6–14. These show that the optimal number of marker genes varies widely from dataset to dataset and method to method.

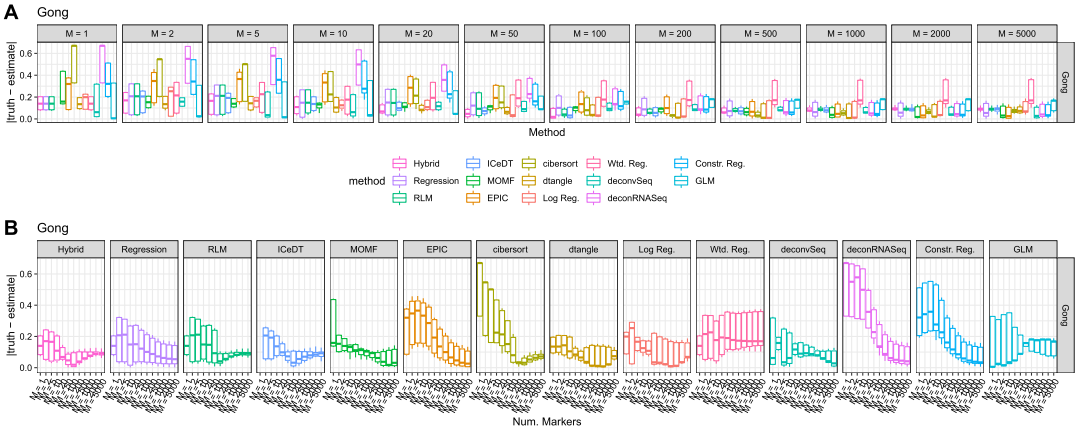

FIG. 5. *Error for methods for the Gong dataset over a varying number of markers* (M). *Error is measured as the absolute value of the truth less the estimate.* (*A*) *displays the plots by number of markers.* (*B*) *displays the exact same data but separating by method.*

Unfortunately, while deconvolution performance is greatly affected by the choice of marker genes, there is no general consensus on an approach to find the optimal number of marker genes for all possible datasets. Nonetheless, modeling using the hybrid-scale approach maintains an error competitive with existing methods and, generally, has estimates that are less sensitive to the number of marker genes. To see this, in Figure 7 we plot a meta-analysis of all methods across all datasets and number of markers. In addition to the methods previously mentioned, we also include a robust regression and weighted regression approach. Each point in this figure is the MAD (median absolute-deviation) error for each dataset for a particular method. A separate subplot is made for each choice of number of markers (per cell type). The boxplots summarize the performance of each method across all the datasets. From this figure we see that the hybrid-scale modeling reduces the grand median absolute error in comparison with other modeling approaches. This is most evident when the number of marker genes (per cell type) is in a middling range between about 20 and 200 (per cell type). In Figure 8 we plot the 1st quartile, median and 3rd quartile of the error across datasets for each method. We vary the number of markers over the x-axis. Again, we see the hybrid approach finds its lowest median error for middling values of number of markers. Indeed,

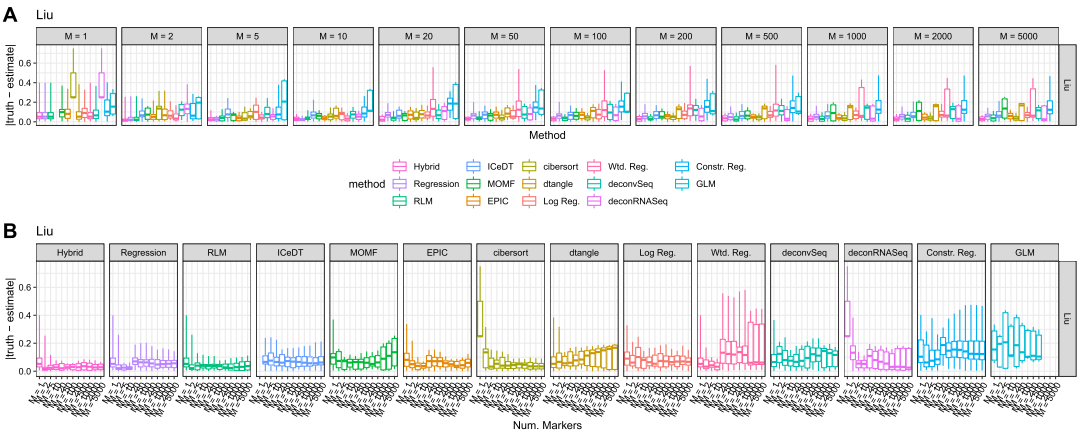

FIG. 6. *Error for methods for the Liu dataset over a varying number of markers* (M). *Error is measured as the absolute value of the truth less the estimate.* (*A*) *displays the plots by number of markers.* (*B*) *displays the exact same data but separating by method.*

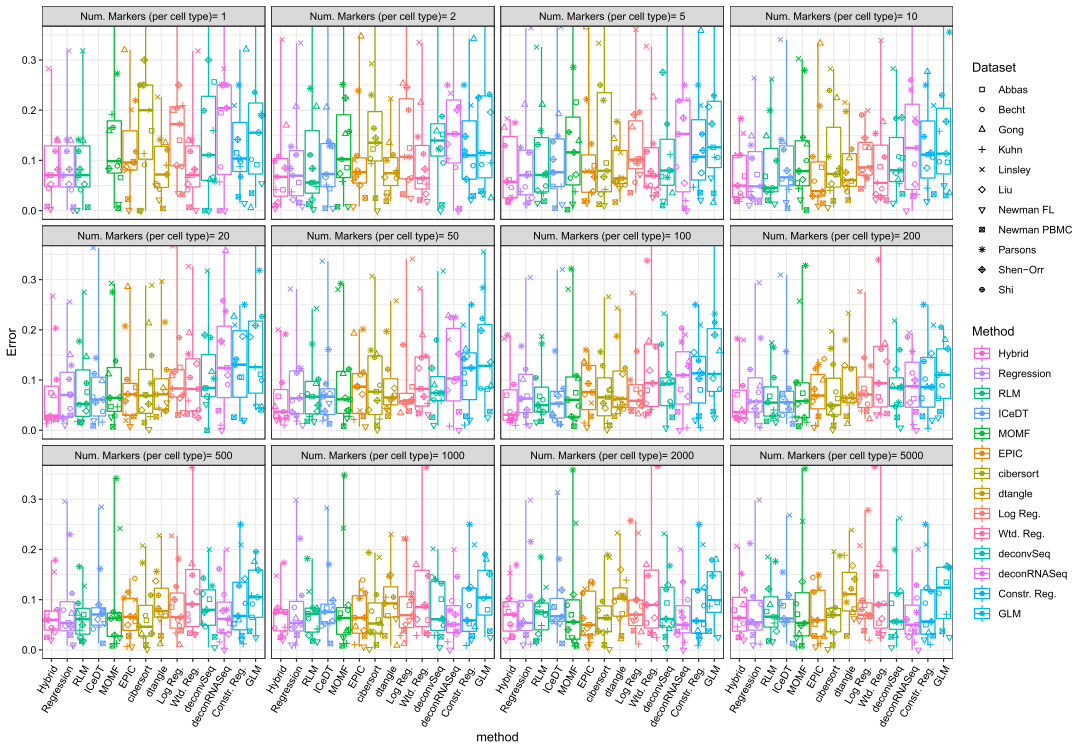

FIG. 7. *MAD error meta-analysis of all methods across all number of markers and datasets.*

the lowest overall error for any method is obtained by this approach in this region. We also similarly see that this approach yields 1st and 3rd quartiles of its error commensurate with the best other approaches.

5. Conclusion. Understanding cell-type heterogeneity among complex biological tissues is a problem with broad and persistent biological interest. Furthermore, an increase in high-quality cell-type reference data from bulk and single-cell sequencing technologies makes cell-type deconvolution an increasingly important tool for the analysis of high-throughput data. Many existing deconvolution approaches estimate cell-type proportions using modified regression approaches, as described in the UDAR framework. However, fitting such a model using either linear-scale or log-scale gene expressions will be suboptimal. Log-transforming gene expressions before fitting under a UDAR model biases the estimates.

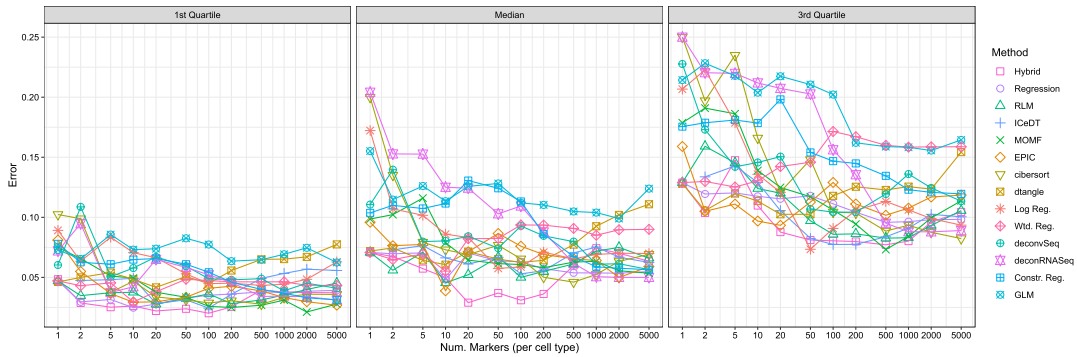

FIG. 8. *Meta-analysis of all error for all methods across all datasets plotted by number of markers.*

However, a regression-like fit using linear-scale gene expressions assumes an unrealistic error model for gene expression measurements. Conversely, a hybrid scale model uses a plausible mean structure while also maintaining reasonable error assumptions. The prototypical hybrid-scale approach we explore in this paper has implications likely applicable to many existing deconvolution methods. We showed that such a model can help improve estimates of cell-type proportions in a robust manner. In simulations the hybrid model reduced estimation variance without introducing a bias and was robust to violations of distributional assumptions. This allowed a broad range of applicability to normalized and unnormalized sequencing data, microarray data and combinations of different technologies. In an analysis of real data, it was shown that cell-type deconvolution is sensitive to choice of marker genes. Unfortunately, this is compounded by the fact that, for real data, there is often no easy way to find an optimal set of marker genes. This is consequential when compared to previous work like dtangle, a method which makes strict assumptions about marker genes, in particular that they are expressed in only one cell type and have zero expression in all other cell types. The hybrid-scale approach does not make the same restrictive assumptions and thus out-performs dtangle when such assumptions are unlikely to hold, for example, when there are many highly-correlated cell types.

More broadly, the model proposed in this paper opens the door to many extensions and generalizations. While we estimate the proportions p in equation (5) using a maximum-likelihood approach, one could combine this model with some of the other insights in the deconvolution literature and fit p using more sophisticated loss functions. For example, one could use L2 penalized losses if there are many highly-correlated cell types (Altboum et al. (2014)). Alternatively, L1 penalties or ε-insensitive losses can be used to induce sparsity which may be beneficial if there are many potential cell types but only a subset that are expected to be present (Fernández et al. (2019), Newman et al. (2015)). Such additions could easily be incorporated into our framework, and, thus, the approach we describe has the potential to be the basis for many new, hybrid-scale, approaches to deconvolution.

6. Software. An implementation of the hybrid approach called `hspe` (Hybrid-Scale Proportion Estimation) and examples of how to use the method can be found online at `gjhunt.github.io`. A `docker` image with code to reproduce all results in this paper may be found at `hub.docker.com/r/gjhunt/hybriddeconv`. Source code for reproducibility can be found in the Supplementary Materials (Hunt and Gagnon-Bartsch (2021)).

Acknowledgments. The authors gratefully acknowledge support from the National Science Foundation (grant no. DMS-1646108).

SUPPLEMENTARY MATERIAL

Supplement to "The role of scale in the estimation of cell-type proportions" (DOI: 10.1214/20-AOAS1395SUPPA; .pdf). A proof of the MLE and figures for simulation and real-data analysis.

Source code for "The role of scale in the estimation of cell-type proportions" (DOI: 10.1214/20-AOAS1395SUPPB; .zip). R source code for reproducibility.

REFERENCES

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

ALTBOUM, Z., STEUERMAN, Y., DAVID, E., BARNETT-ITZHAKI, Z., VALADARSKY, L., KEREN-SHAUL, H., MENINGHER, T., MENDELSON, E., MANDELBOIM, M. et al. (2014). Digital cell quantification identifies global immune cell dynamics during influenza infection. *Mol. Syst. Biol.* **10** 1–14. https://doi.org/10.1002/msb.134947

AYANA, R., SINGH, S. and PATI, S. (2018). Deconvolution of human brain cell type transcriptomes unraveled microglia-specific potential biomarkers. *Front. Neurology* **9** 266. https://doi.org/10.3389/fneur.2018.00266

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation **3** 993–1022.

BOUQUET, J., SOLOSKI, M. J., SWEI, A., CHEADLE, C., FEDERMAN, S., BILLAUD, J.-N. and REBMAN, A. W. (2016). Longitudinal transcriptome analysis reveals a sustained differential gene expression signature in patients treated for acute lyme disease **7** 1–11. https://doi.org/10.1128/mBio.00100-16.Editor

CAPURRO, A., BODEA, L. G., SCHAEFER, P., LUTHI-CARTER, R. and PERREAU, V. M. (2015). Computational deconvolution of genome wide expression data from Parkinson's and Huntington's disease brain tissues using population-specific expression analysis. *Front. Neurosci.* **9** 1–12. https://doi.org/10.3389/fnins.2014.00441

CONESA, A., MADRIGAL, P., TARAZONA, S., GOMEZ-CABRERO, D., CERVERA, A., MCPHERSON, A., SZCZEŚNIAK, M. W., GAFFNEY, D. J., ELO, L. L. et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17** 13. https://doi.org/10.1186/s13059-016-0881-8

DU, R., CAREY, V. and WEISS, S. T. (2019). DeconvSeq: Deconvolution of cell mixture distribution in sequencing data. *Bioinformatics* **35** 5095–5102. https://doi.org/10.1093/bioinformatics/btz444

EDGAR, R. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30** 207–210. https://doi.org/10.1093/nar/30.1.207

FERNÁNDEZ, E., MAHMOUD, Y., VEIGAS, F., ROCHA, D., BALZARINI, M., LUJAN, H., RABINOVICH, G. and GIROTTI, M. R. (2019). MIXTURE: An improved algorithm for immune tumor microenvironment estimation based on gene expression data. *BioRxiv* 726562. https://doi.org/10.1101/726562

FINOTELLO, F., MAYER, C., PLATTNER, C., LASCHOBER, G., RIEDER, D., HACKL, H., KROGSDAM, A., LONCOVA, Z., POSCH, W. et al. (2019). Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Gen. Med.* **11** 34. https://doi.org/10.1186/s13073-019-0638-6

GAUJOUX, R. (2013). An introduction to gene expression deconvolution and the CellMix package. 1–45.

GEORGE, E. L. and PANOS, A. (2007). Does a high WBC count always signal infection? *Nursing* **37** 56hn15–56hn16. https://doi.org/10.1097/01.NURSE.0000268785.73612.5c

GONG, T. and SZUSTAKOWSKI, J. D. (2013). DeconRNASeq: A statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-seq data. *Bioinformatics* **29** 1083–1085. https://doi.org/10.1093/bioinformatics/btt090

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

HAGENAUER, M. H., LI, J. Z., WALSH, D. M., VAWTER, M. P., THOMPSON, R. C., TURNER, C. A., BUNNEY, W. E., MYERS, R. M., BARCHAS, J. D. et al. (2016). Inference of cell type composition from human brain transcriptomic datasets illuminates the effects of age, manner of death, dissection, and psychiatric diagnosis. *BioRxiv*.

HARDIN, J. and WILSON, J. (2009). A note on oligonucleotide expression values not being normally distributed. *Biostatistics* **10** 446–450. https://doi.org/10.1093/biostatistics/kxp003

HUNT, G. J. and GAGNON-BARTSCH, J. A. (2021). Supplement to "The Role of Scale in the Estimation of Cell-type Proportions." https://doi.org/10.1214/20-AOAS1395SUPPA, https://doi.org/10.1214/20-AOAS1395SUPPB

HUNT, G. J., FREYTAG, S., BAHLO, M. and GAGNON-BARTSCH, J. A. (2019). dtangle: Accurate and robust cell type deconvolution. *Bioinformatics* **35** 2093–2099. https://doi.org/10.1093/bioinformatics/bty926

IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. and SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** 249–264. https://doi.org/10.1093/biostatistics/4.2.249

SUN, X., SUN, S. and YANG, S. (2019). An efficient and flexible method for deconvoluting bulk RNA-seq data with single-cell RNA-seq data. *Cells* **8** 1161. https://doi.org/10.3390/cells8101161

LIU, R., HOLIK, A. Z., SU, S., JANSZ, N., CHEN, K., LEONG, S., BLEWITT, M. E., SMYTH, G. K. and RITCHIE, M. E. (2015). Why weight ? Modelling sample and observational level variability improves power in RNA-seq analyses **43**. https://doi.org/10.1093/nar/gkv412

LU, P., NAKORCHEVSKIY, A. and MARCOTTE, E. M. (2003). Expression deconvolution: A reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proc. Natl. Acad. Sci. USA* **100** 10370–10375. https://doi.org/10.1073/pnas.1832361100

MOHAMMADI, S., ZUCKERMAN, N., GOLDSMITH, A. and GRAMA, A. (2015). A Critical Survey of Deconvolution Methods for Separating cell-types in Complex Tissues. arXiv 1–20.

NEWMAN, A. M., LONG LIU, C., GREEN, M. R., GENTLES, A. J., FENG, W., XU, Y., HOANG, C. D., DIEHN, M. and ALIZADEH, A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12** 193–201. https://doi.org/10.1016/j.molmed.2014.11.008.Mitochondria

PARSONS, J., MUNRO, S., PINE, P. S., MCDANIEL, J., MEHAFFEY, M. and SALIT, M. (2015). Using mixtures of biological samples as process controls for RNA-sequencing experiments. *BMC Genomics* 1–13. https://doi.org/10.1186/s12864-015-1912-7

QIAO, W., QUON, G., CSASZAR, E., YU, M., MORRIS, Q. and ZANDSTRA, P. W. (2012). PERT: A method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput. Biol.* **8**. https://doi.org/10.1371/journal.pcbi.1002838

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

TU, Y., STOLOVITZKY, G. and KLEIN, U. (2002). Quantitative noise analysis for gene expression microarray experiments. *Proc. Natl. Acad. Sci. USA* **99** 14031–14036. MR1944414 https://doi.org/10.1073/pnas.222164199

WANG, M., MASTER, S. R. and CHODOSH, L. A. (2006). Computational expression deconvolution in a complex mammalian organ. *BMC Bioinform.* **7** 328. https://doi.org/10.1186/1471-2105-7-328

WANG, N., HOFFMAN, E. P., CHEN, L., CHEN, L., ZHANG, Z., LIU, C., YU, G., HERRINGTON, D. M., CLARKE, R. et al. (2016). Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Sci. Rep.* **6** 18909. https://doi.org/10.1038/srep18909

WANG, Z., CAO, S., MORRIS, J. S., AHN, J., LIU, R., TYEKUCHEVA, S., GAO, F., LI, B., LU, W. et al. (2018). Transcriptome deconvolution of heterogeneous tumor samples with immune infiltration. *IScience* **9** 451–460. https://doi.org/10.1016/j.isci.2018.10.028

WENG, L., DAI, H., ZHAN, Y., HE, Y., STEPANIANTS, S. B. and BASSETT, D. E. (2006). Rosetta error model for gene expression analysis. *Bioinformatics* **22** 1111–1121. https://doi.org/10.1093/bioinformatics/btl045

WILSON, D. R., JIN, C., IBRAHIM, J. G. and SUN, W. (2020). ICeD-T provides accurate estimates of immune cell abundance in tumor samples by allowing for Aberrant gene expression patterns. *J. Amer. Statist. Assoc.* **115** 1055–1065. MR4143449 https://doi.org/10.1080/01621459.2019.1654874

ZWIENER, I., FRISCH, B. and BINDER, H. (2014). Transforming RNA-seq data to improve the performance of prognostic gene signatures. *PLoS ONE* **9** e85150. https://doi.org/10.1371/journal.pone.0085150

SUPPLEMENTARY MATERIALS: THE ROLE OF SCALE IN THE ESTIMATION OF CELL-TYPE PROPORTIONS

GREGORY J. HUNT[1,*], JOHANN A. GAGNON-BARTSCH[2]
[1] *DEPARTMENT OF MATHEMATICS, WILLIAM & MARY*
[2] *DEPARTMENT OF STATISTICS, UNIVERSITY OF MICHIGAN*
GHUNT@WM.EDU

1. Maximum Likelihood Estimator for Hybrid Model

We have claimed that our fitting approach described in Algorithm 2 gives the MLE for our model. We will now show that. First we will show that maximizing the log-likelihood is equivalent to minimizing $S^2(p)$ in Algorithm 2. Consider the joint log-likelihood ℓ of the parameters p, σ^2, and θ. This log-likelihood is (up to an additive constant)

$$\begin{aligned}\ell(p,\sigma^2,\theta) &= -\frac{N}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(\log(Y_n) - \theta - \log\left(\sum_{k=1}^{K} p_k R_{nk}\right)\right)^2 \\ &= -\frac{N}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(\lambda_n(p) - \theta\right)^2\end{aligned} \tag{1}$$

where $\lambda_n(p) \stackrel{def}{=} \log(Y_n) - \log\left(\sum_{k=1}^{K} p_k R_{nk}\right)$. The partial derivative with respect to θ is

$$\frac{\partial \ell}{\partial \theta} = \frac{1}{\sigma^2}\sum_{n=1}^{N}\left(\lambda_n(p) - \theta\right) \tag{2}$$

and setting this equal to zero we get the maximizing value of θ is

$$\theta^*(p) = \bar{\lambda}(p) \stackrel{def}{=} \frac{1}{N}\sum_{n=1}^{N}\lambda_n(p). \tag{3}$$

Furthermore, the partial derivative with respect to σ^2 is

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=1}^{N}\left(\lambda_n(p) - \theta\right)^2 \tag{4}$$

and setting this equal to zero we get the maximizing value for σ^2 to be

$$\sigma^{2*}(p,\theta) = \frac{1}{N}\sum_{n=1}^{N}\left(\lambda_n(p) - \theta\right)^2. \tag{5}$$

We can then form the profile likelihood for p as

$$\ell^*(p) = \ell\left(p, \sigma^{2*}(p,\theta^*(p)), \theta^*(p)\right) = -\frac{N}{2}\log(\sigma^{2*}(p,\theta^*(p))) - \frac{N}{2}. \tag{6}$$

Thus we can find the MLE for p by maximizing $\ell^*(p)$ or equivalently minimizing

$$\sigma^{2*}(p, \theta^*(p)) = \frac{1}{N} \sum_{n=1}^{N} \left(\lambda_n(p) - \bar{\lambda}(p)\right)^2 \tag{7}$$

which is precisely the sample variance of the $\lambda_n(p)$ defined as $S^2(p)$ in Algorithm 2.

Second, we will show that minimizing $S^2(p)$ over $[0,1]^K$ and then re-normalizing is equivalent to minimizing $S^2(p)$ over Δ_{K-1}. Now if

$$p^* = \arg \min_{p \in [0,1]^K} S^2(p) \tag{8}$$

then, letting $T^* = \sum_{t=1}^{K} p_t^*$, Algorithm 2 defines $\hat{p} = p^*/T^*$. To see that $\hat{p}$ is the MLE, note that for a constant $c \in \mathbb{R}$,

$$\lambda_n(cp) = \log(Y_n) - \log\left(\sum_{k=1}^{K} cpR_{nk}\right) = \log(Y_n) - \log\left(\sum_{k=1}^{K} pR_{nk}\right) - \log(c) = \lambda_n(p) - \log(c).$$

and thus $S^2(cp) = S^2(p)$ as $S^2(p)$ is the variance of the $\lambda_n(p)$ and $S^2(cp)$ is the variance of the $\lambda_n(cp) = \lambda_n(p) - \log(c)$ and the two are equivalent since subtracting a constant $\log(c)$ does not change the sample variance S^2. Thus, $S^2(\hat{p}) = S^2\left(p^*/T^*\right) = S^2(p^*)$. Then since $S^2(\hat{p}) = S^2(p^*) \leq S^2(p)$ for all $p \in \Pi$ and $\Delta_{K-1} \subseteq \Pi$ then $S^2(\hat{p}) = S^2(p^*) \leq S^2(p)$ for all $p \in \Delta_{K-1}$ and so $\hat{p}$ minimizes S^2 over Δ_{K-1}. Equivalently, $\hat{p}$ is the MLE of p over Δ_{K-1}.

Name	Citation	Cell Types	Species
Shi	MAQC (2006)	2, universal, brain	human
Gong	Gong *et al.* (2011)	2, blood, breast	human
Shen-Orr	Shen-Orr *et al.* (2010)	3, liver, brain, lung	rat
Abbas	Abbas *et al.* (2009)	4, leukocytes	human
Becht	Becht *et al.* (2016)	6, colorectal carcinoma, leukocytes	human
Kuhn	Kuhn *et al.* (2011)	4, brain	rat
Newman FL	Newman *et al.* (2015)	12, leukocytes	human
Newman PBMC	Newman *et al.* (2015)	22, leukocytes	human
Parsons	Parsons *et al.* (2015)	3, brain, liver, muscle	human
Liu	Liu *et al.* (2015)	2, adenocarcinoma	human
Linsley	Linsley *et al.* (2014)	3, lymphocytes,monocytes, neutrophils	human

TABLE 1. Eleven benchmark deconvolution datasets.

2. BENCHMARK DATASETS AND SUPPLEMENTARY FIGURES

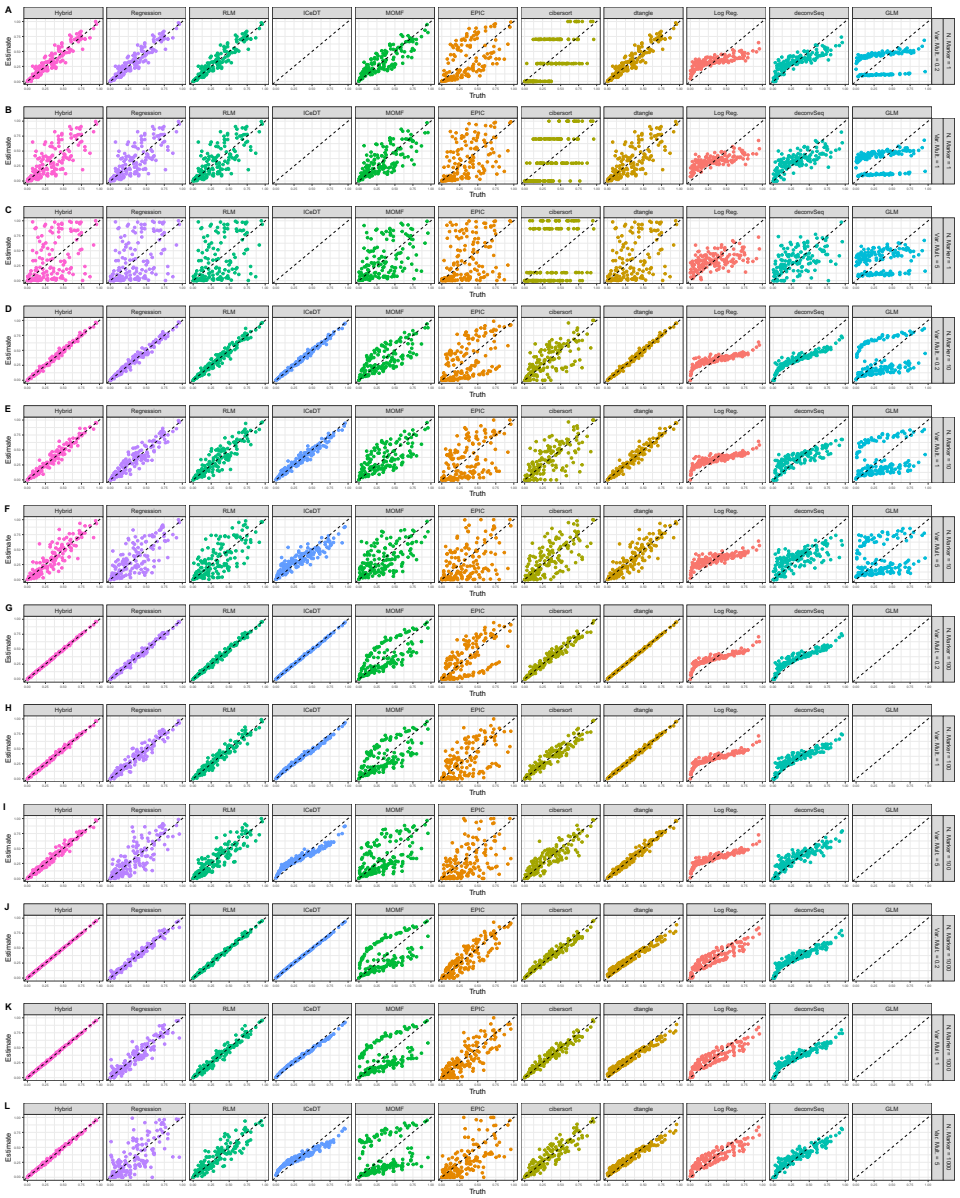

FIGURE 1. Similar to Figure 2 but over a wider range of simulation settings.

FIGURE 2. Similar to Figure 2 but over a wider range of simulation settings and including more methods. The reference for this data is 22 white blood cell types and thus is a scenario with many closely-related cell types.

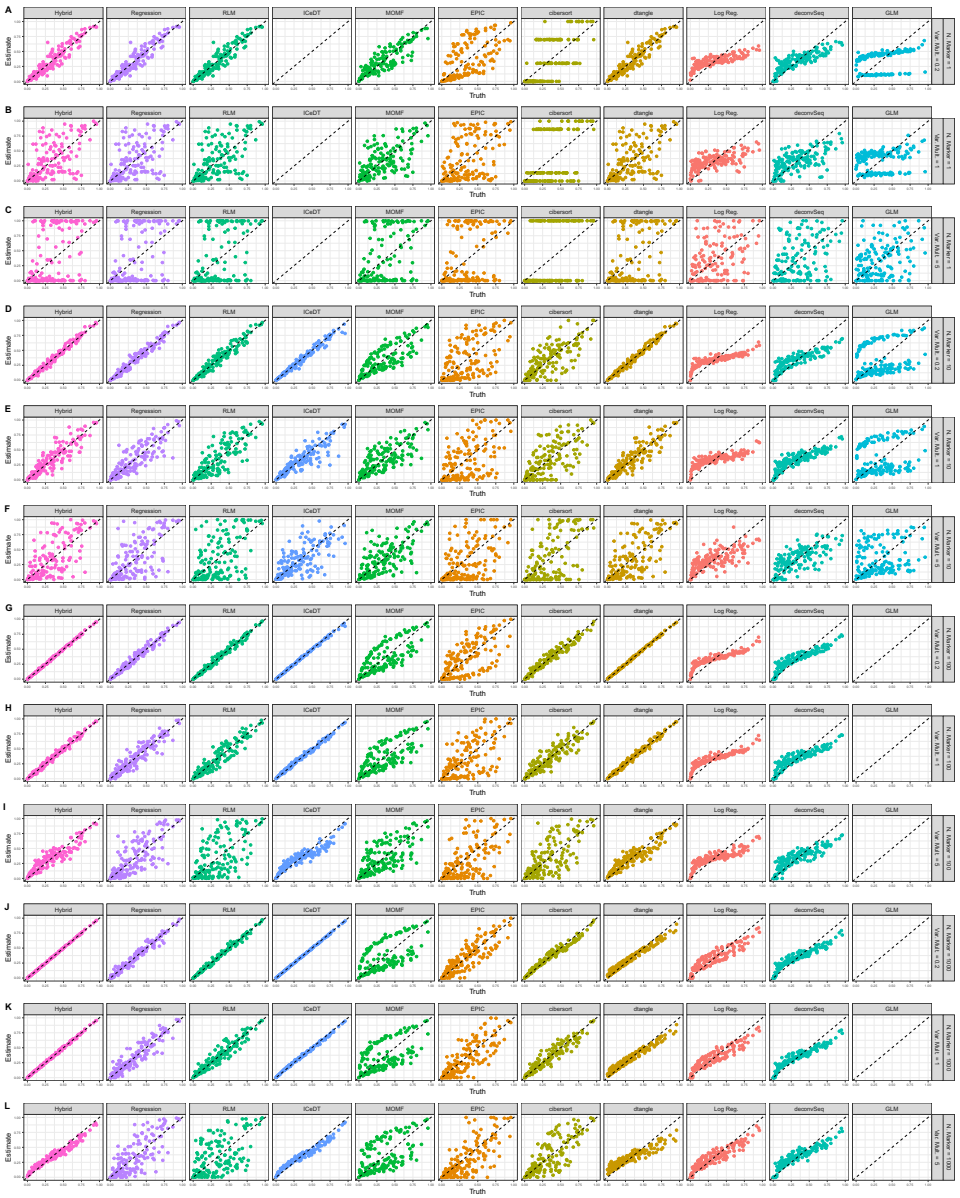

FIGURE 3. Similar to Figure 3 but over a wider range of simulation settings.

FIGURE 4. Similar to Figure 3 but over a wider range of simulation settings and including more methods. The reference for this data is 22 white blood cell types and thus is a scenario with many closely-related cell types.

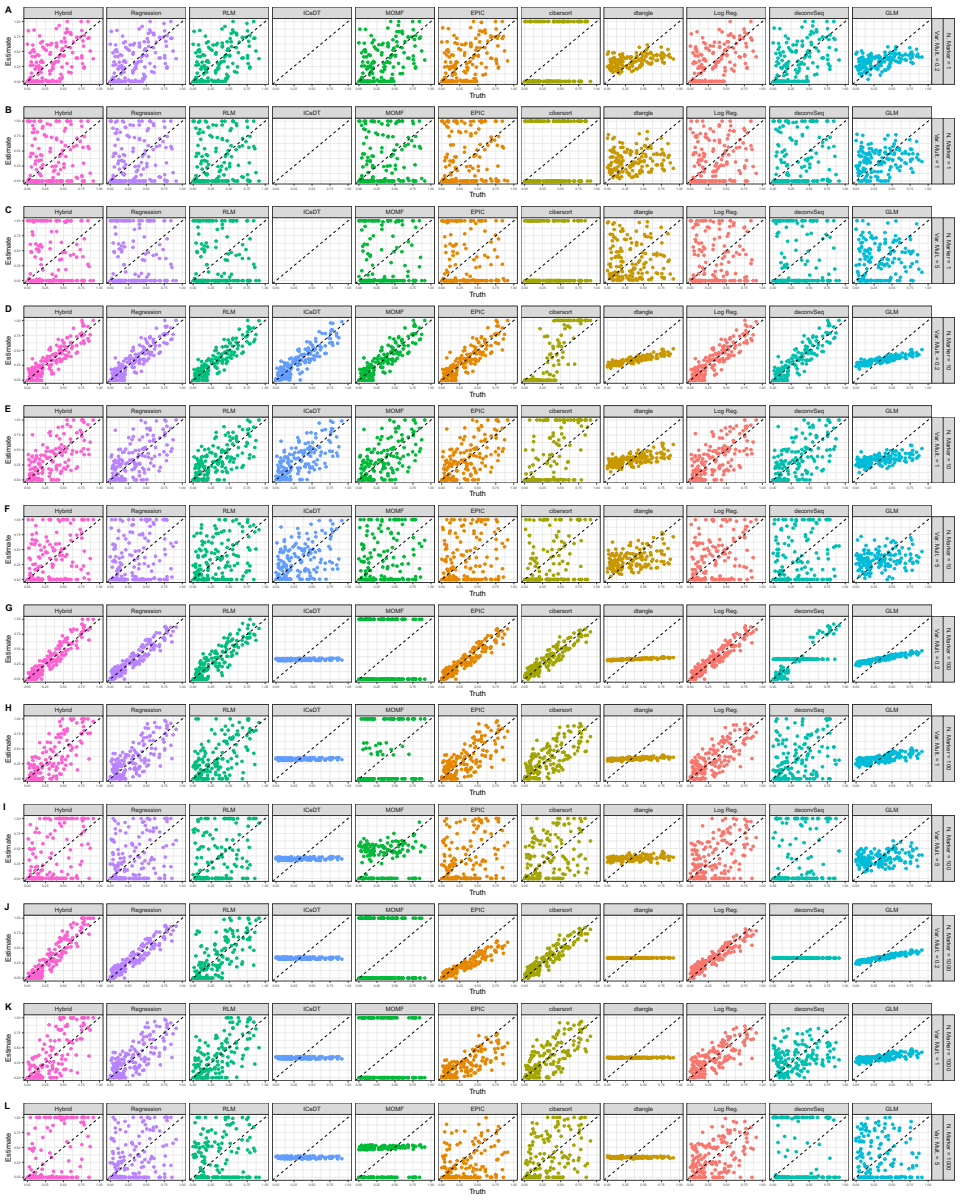

FIGURE 5. Similar to Figure 4 comparing all methods over all simulation settings.

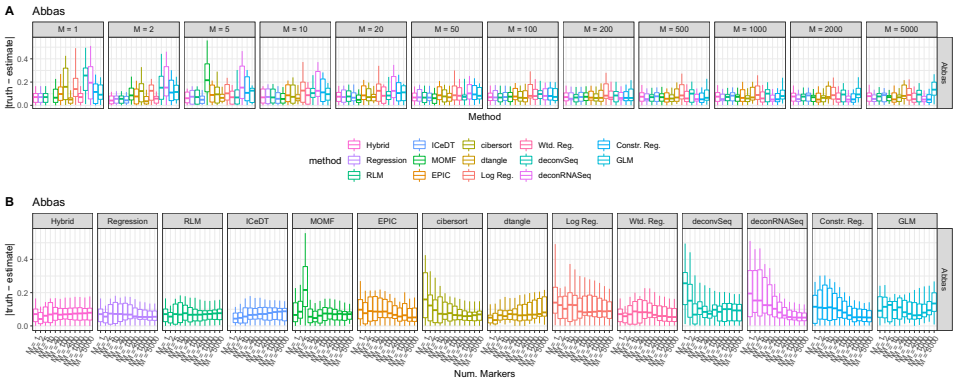

FIGURE 6. Error for methods for the Abbas dataset over a varying number of markers (M). Error is measured as the absolute value of the truth less the estimate. (A) displays the plots by number of markers. (B) displays the exact same data but separating by method.

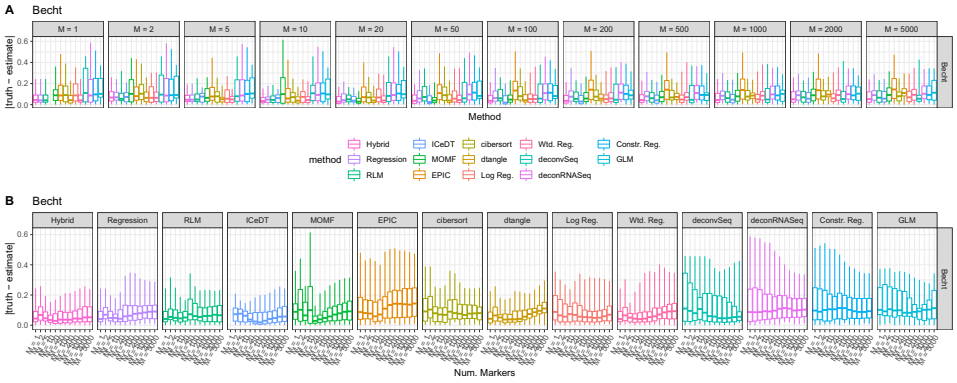

FIGURE 7. Similar to Figure 6 but for the Becht dataset.

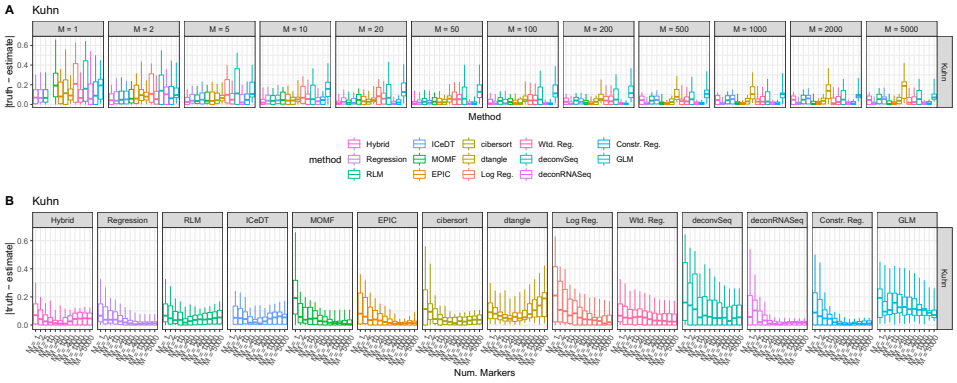

FIGURE 8. Similar to Figure 6 but for the Kuhn dataset.

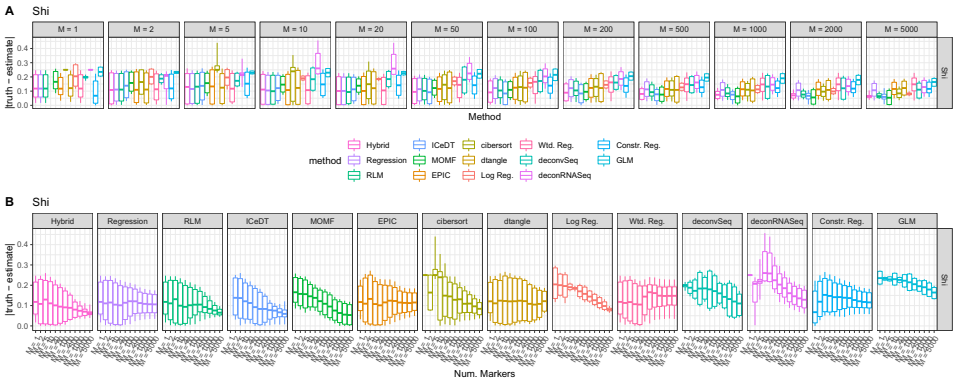

FIGURE 9. Similar to Figure 6 but for the Shi dataset.

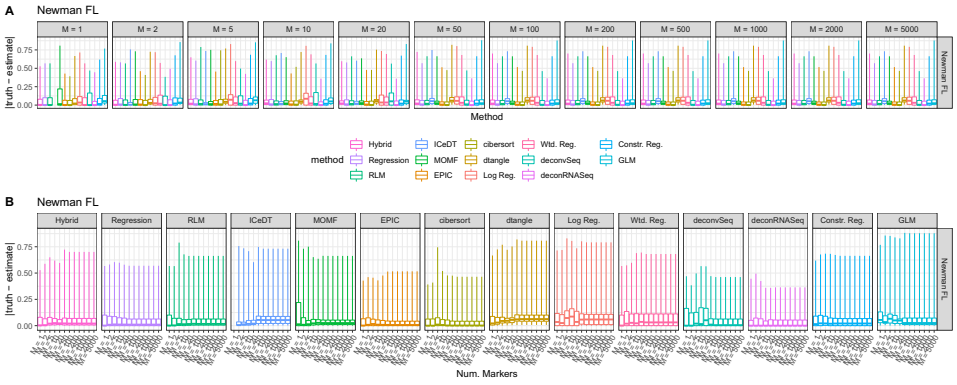

FIGURE 10. Similar to Figure 6 but for the Newman FL dataset.

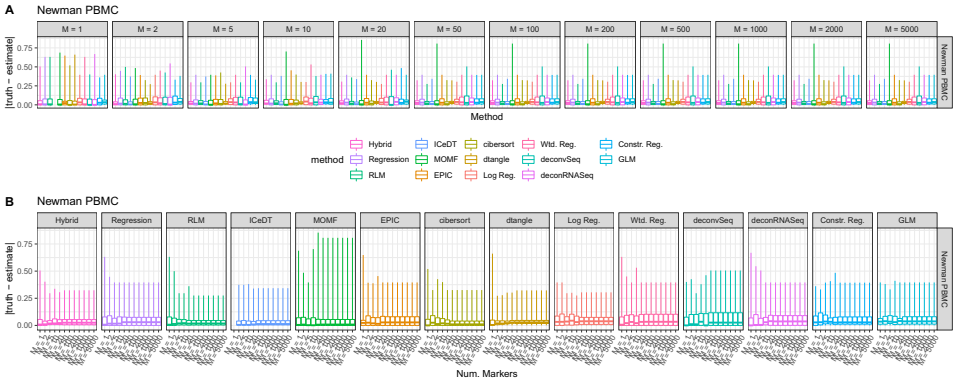

FIGURE 11. Similar to Figure 6 but for the Newman PBMC dataset.

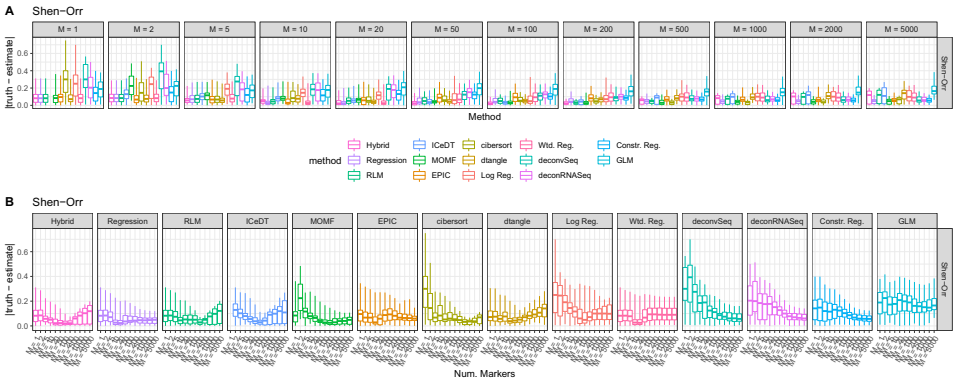

FIGURE 12. Similar to Figure 6 but for the Shen-Orr dataset.

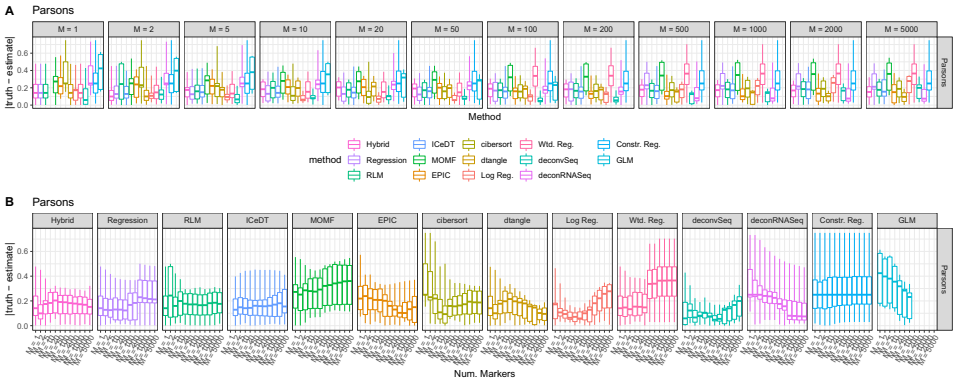

FIGURE 13. Similar to Figure 6 but for the Parsons dataset.

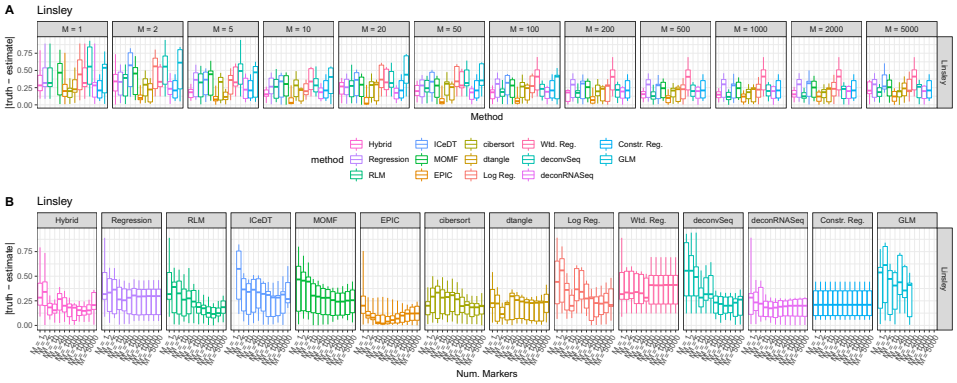

FIGURE 14. Similar to Figure 6 but for the Linsley dataset.

3. Modification of ICeD-T Algorithm

We made 3 modifications to the ICeD-T code. We added `na.rm=TRUE` to two calls to the quantile function and commented out two lines where an exception was thrown when the likelihood decreased. These do not alter the behavior of ICeD-T much but stop it from failing in many scenarios. The file diff is as follows:

```
43c43
<     Q3val = quantile(abs(resid), probs = c(0.75))
---
>     Q3val = quantile(abs(resid), probs = c(0.75), na.rm = TRUE)
308,310c308,309
<         if (logLik_1 < logLik_0 - 0.001) {
<             stop("log likelihood decreased during the EM algorithm.")
<         }
---
>         # if(logLik_1 < logLik_0 - 0.001){ stop('log likelihood decreased during the EM
>         # algorithm.') }
607c606
<         quants = quantile(var_wgt, prob = c(0.15, 0.85))
---
>         quants = quantile(var_wgt, prob = c(0.15, 0.85), na.rm = TRUE)
```

References

[1] Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z., and Clark, H. F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS ONE*, **4**(7).

[2] Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P., Sautès-Fridman, C., Fridman, W. H., and de Reyniès, A. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome biology*, **17**(1), 218.

[3] Gong, T., Hartmann, N., Kohane, I. S., Brinkmann, V., Staedtler, F., Letzkus, M., Bongiovanni, S., and Szustakowski, J. D. (2011). Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS ONE*, **6**(11).

[4] Kuhn, A., Thu, D., Waldvogel, H. J., Faull, R. L. M., and Luthi-Carter, R. (2011). Population-specific expression analysis (PSEA) reveals molecular changepcts re-strict to markers, s in diseased brain. *Nature methods*, **8**(11), 945–7.

[5] 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.

[6] Liu, R., Holik, A. Z., Su, S., Jansz, N., Chen, K., Leong, S., Blewitt, M. E., Smyth, G. K., and Ritchie, M. E. (2015). Why weight ? Modelling sample and observational level variability improves power in RNA-seq analyses. **43**(15).

[7] MAQC (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. **24**(9), 1151–1161.

[8] Newman, A. M., Chih Long Liu, Michael R. Green, Andrew J. Gentles, W. F., Yue Xu, C. D. H., Diehn, M., and Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods.*, **12**(5), 193–201.

[9] Parsons, J., Munro, S., Pine, P. S., Mcdaniel, J., Mehaffey, M., and Salit, M. (2015). Using mixtures of biological samples as process controls for RNA-sequencing experiments. *BMC Genomics*, pages 1–13.

[10] Shen-Orr, S. S., Tibshirani, R., Khatri, P., Bodian, D. L., Staedtler, F., Perry, N. M., Hastie, T., Sarwal, M. M., Davis, M. M., and Butte, A. J. (2010). Cell type-specific gene expression differences in complex tissues. *Nat Methods*, **7**(4), 287–289.