Bioinformatics, 38(19), 2022, 4530–4536
https://doi.org/10.1093/bioinformatics/btac563
Advance Access Publication Date: 18 August 2022
Original Paper

Gene expression

Omnibus and robust deconvolution scheme for bulk RNA sequencing data integrating multiple single-cell reference sets and prior biological knowledge

Chixiang Chen [1,2,*], Yuk Yee Leung [3,4], Matei Ionita[3,4], Li-San Wang[3,4,*] and Mingyao Li[5,*]

[1]Department of Epidemiology and Public Health, Division of Biostatistics and Bioinformatics, University of Maryland School of Medicine, Baltimore, MD 21201, USA, [2]Department of Neurosurgery, University of Maryland School of Medicine, Baltimore, MD 21201, USA, [3]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA, [4]Department of Pathology and Laboratory Medicine, Penn Neurodegeneration Genomics Center, Philadelphia, PA 19104, USA and [5]Department of Biostatistics Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA

*To whom correspondence should be addressed.
Associate Editor: Anthony Mathelier

Received on January 19, 2022; revised on July 17, 2022; editorial decision on August 8, 2022; accepted on August 17, 2022

Abstract

Motivation: Cell-type deconvolution of bulk tissue RNA sequencing (RNA-seq) data is an important step toward understanding the variations in cell-type composition among disease conditions. Owing to recent advances in single-cell RNA sequencing (scRNA-seq) and the availability of large amounts of bulk RNA-seq data in disease-relevant tissues, various deconvolution methods have been developed. However, the performance of existing methods heavily relies on the quality of information provided by external data sources, such as the selection of scRNA-seq data as a reference and prior biological information.

Results: We present the Integrated and Robust Deconvolution (InteRD) algorithm to infer cell-type proportions from target bulk RNA-seq data. Owing to the innovative use of penalized regression with a new evaluation criterion for deconvolution, InteRD has three primary advantages. First, it is able to effectively integrate deconvolution results from multiple scRNA-seq datasets. Second, InteRD calibrates estimates from reference-based deconvolution by taking into account extra biological information as priors. Third, the proposed algorithm is robust to inaccurate external information imposed in the deconvolution system. Extensive numerical evaluations and real-data applications demonstrate that InteRD yields more accurate and robust cell-type proportion estimates that agree well with known biology.

Availability and implementation: The proposed InteRD framework is implemented in R and the package is available at https://cran.r-project.org/web/packages/InteRD/index.html.

Contact: chixiang.chen@som.umaryland.edu or lswang@pennmedicine.upenn.edu or mingyao@pennmedicine.upenn.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Solid tissue is composed of diverse cell types, each of which has a distinct biological function. Knowledge of cell-type composition and the corresponding cell-type proportions is important because certain cell types are more vulnerable to diseases than others. Failure to account for variation in cell-type composition in bulk tissue RNA sequencing (RNA-seq) analysis may lead to false detection of differentially expressed genes, in which the apparent differential expression is simply driven by variation in cell-type composition among samples between conditions. Recent advances in single-cell RNA sequencing (scRNA-seq) have enabled the characterization of cell-type-specific gene expression (Hrdlickova *et al.*, 2017). However, scRNA-seq is still costly, which has limited its applications in clinical studies that involve large numbers of samples. Furthermore, owing to the potential bias induced by cell dissociation, cell-type proportions estimated from scRNA-seq data may not reflect the true cell-type proportions in the original intact bulk tissue. A possible way to overcome these limitations is to perform cell-type deconvolution in bulk RNA-seq samples by integrating cell-type-specific gene expression information provided by scRNA-seq (Chen *et al.*, 2018; Kuhn *et al.*, 2011).

Substantial methodologies on cell-type deconvolution from bulk RNA-seq data have emerged in recent years. By taking advantage of scRNA-seq data, current state-of-the-art deconvolution algorithms, named reference-based (RB) approaches, rely almost exclusively on cell-type-specific gene expression profiles (GEPs) provided by the scRNA-seq reference (Newman *et al.*, 2015, 2019; Wang *et al.*, 2019). However, the applications of RB approaches are subject to the use of reference panels, which may not fully reflect the unobserved cell-type-specific GEPs in the bulk RNA-seq samples. The correction for batch effects between datasets could be an alternative, but its performance on improving deconvolution is at least not clear (Dong *et al.*, 2021). Moreover, as increasing number of single-cell data become publicly available, methods that only use one scRNA-seq reference at a time have become more questionable. Recently, innovative methods including SCDC-ENSEMBLE (Dong *et al.*, 2021) and SCADEN (Menden *et al.*, 2020) have been proposed to handle multiple scRNA-seq reference sets. However, the former method uses distinct GEPs from each scRNA-seq dataset to replace the underlying GEPs in bulk data and thus still yields vulnerable estimates; the latter approach pools multiple reference sets into a deep learning machine, which could also be questionable without discrimination of the quality and heterogeneity across the scRNA-seq data.

To alleviate the above-mentioned limitations of RB deconvolution methods, reference-free methods have been developed (Houseman *et al.*, 2014; Li *et al.*, 2020; Repsilber *et al.*, 2010). These methods typically use factorization-based approaches such as non-negative matrix factorization to infer cell-type proportions. However, such techniques may suffer from low accuracy and have difficulty in assigning cell-type labels owing to a lack of supporting information on cell-type-specific GEPs. To enhance the reference-free strategy, Li *et al.* (2020) recently developed TOAST/+P, a partial reference-free approach for gene expression microarray data by incorporating additional biological information as priors (e.g. population-level mean proportions and standard deviations for each cell type) into the model. Although it has shown better performance than competing algorithms, TOAST/+P implicitly assumes that the prior biological information is accurate, which is hard to achieve and validate in practice. In addition, the estimated cell-type proportions from TOAST/+P shrink toward the population means and hence may fail to reveal variations in cell-type proportions among subjects. Failure to utilize cell-type-specific GEPs from scRNA-seq data in TOAST/+P further limits its practical applicability.

To address these challenges, we introduced a unified computational framework, named **Inte**grated and **R**obust **D**econvolution (InteRD), to infer cell-type proportions from bulk RNA-seq data. The advantages of the proposed framework are 3-fold. First, InteRD is able to integrate deconvolution results from multiple scRNA-seq datasets without assuming that GEPs in different reference sets are similar to those in the underlying bulk tissue. Second, InteRD calibrates the RB estimates by incorporating a reference-free approach and taking into account prior biological knowledge, such as population-level mean proportions. This boosts the deconvolution performance by incorporating more information into the deconvolution system. Last, the proposed algorithm is equipped with a data-driven mechanism of self-control designed to be robust to introduction of inaccurate information into the system, such as less representative scRNA-seq datasets and/or inaccurate prior knowledge of population-level mean proportions. As a result, InteRD unifies reference-free, RB and partial-reference-free methods into one framework to conduct more powerful and robust deconvolution; it is the first method of this type to be reported in the literature (Supplementary Table S3). By analyzing benchmark data, we demonstrated that InteRD successfully captured the most appropriate reference or combination of multiple reference sets and led to more robust and precise estimates. Furthermore, by analyzing a human pancreas islet bulk RNA-seq dataset that included both healthy and type 2 diabetes (T2D) samples, we demonstrated the effectiveness of InteRD in recovering the functional profiles of beta cells during T2D progression. We also used InteRD to decompose bulk tissue data from the dorsolateral prefrontal cortex of human subjects. Decreased proportions of neurons and increased proportions of microglia were detected in donors with Alzheimer's disease (AD) compared with controls.

2 Motivating data and method overview

For the application involving human pancreatic islets, the target bulk RNA-seq dataset was taken from the study of Fadista *et al.* (2014) and included a total of 77 study subjects collected from donors with and without T2D. In addition to the target data, there were three single-cell RNA-seq datasets available, from 3 healthy adult donors (Baron *et al.*, 2016), 6 healthy and 4 T2D adult donors (Segerstolpe *et al.*, 2016) and 12 healthy and 6 T2D adult donors (Xin *et al.*, 2016). For the application involving human dorsolateral prefrontal cortex, the target bulk RNA-seq data were taken from the study of De Jager *et al.* (2018) and consisted of postmortem brains from 155 AD and 86 control donors. Besides, two single-nucleus RNA-seq (snRNA-seq) datasets were available for middle frontal neocortex brain tissue Nguyen *et al.* (2020) and prefrontal cortex (Mathys *et al.*, 2019). The former snRNA-seq dataset contained data from donors with three different degrees of neurofibrillary degeneration, whereas the latter was from donors with AD and controls. The availability of multiple scRNA-seq reference datasets involving disparate disease conditions is common for many tissues. InteRD first integrates proportion estimates based on multiple candidate reference sets (Fig. 1A). Then, it aggregates prior biological knowledge with integrated RB estimates to recover more precise and robust estimates (Fig. 1B and C).

3 Materials and methods

3.1 Notation and RB deconvolution

Let $\mathbf{Y}$ be the $N \times J$ gene expression matrix in the count scale for bulk RNA-seq data with N samples and J genes, and let $\mathbf{B}^i$ be the $J \times K$ cell-type-specific gene expression matrix for subject i with columns representing the average gene expression levels for the K cell types. Furthermore, we use $\mathbf{P}$ to denote the $N \times K$ cell-type-proportion matrix and m_i to denote the total number of cells for subject i. Note that all vectors and matrices are displayed as bold symbols, all vectors are displayed in column form; and $\mathbf{A}'$ indicates the operation of transposition for any vector or matrix $\mathbf{A}$. As the gene expression count in bulk RNA-seq can be considered as an aggregation of expression over single cells (Dong *et al.*, 2021; Wang *et al.*, 2019), the expression levels in the target bulk data for subject i can be written as:

$$\mathbf{Y}_{i\cdot} = \sum_{k=1}^{K} \sum_{w \in \mathbf{C}_i^k} \mathbf{Y}_{i\cdot}^w = m_i \times \mathbf{B}^i \times \mathbf{P}_{i\cdot}, \tag{1}$$

where $\mathbf{Y}_{i\cdot}^w$ is a vector containing expression levels of genes for cell w, $\mathbf{C}_i^k$ is the set of cell indices for cell type k in subject i and $\mathbf{P}_{i\cdot}$ is a vector containing cell-type proportions for subject i. As both m_i and $\mathbf{B}^i$ are unobserved, most state-of-the-art RB deconvolution algorithms use a $K \times J$ matrix containing mean cell-type-specific GEPs from single-cell data to represent $m_i\mathbf{B}^i$, denoted by $\mathbf{M}$. Thus, given the GEPs, the gene expression matrix $\mathbf{Y}$ can be approximated and cell-type proportions $\mathbf{P}$ can be recovered based on machine learning or regression by $\mathbf{Y} \approx \mathbf{P} \times \mathbf{M}$. After library size normalization for each subject, the cell-type proportions $\mathbf{P}$ can be recovered by the widely applied non-negative least square (NNLS) or machine learning techniques (Newman *et al.*, 2019; Wang *et al.*, 2019). More discussions about parametrically modeling count data are referred to Section 6.

3.2 Step 1: RB integration

State-of-the-art algorithms rely heavily on the quality of reference panels, most of which only use one reference at a time. This section introduces the first step of InteRD, which involves integrating results from multiple reference panels. Suppose there are R single-cell

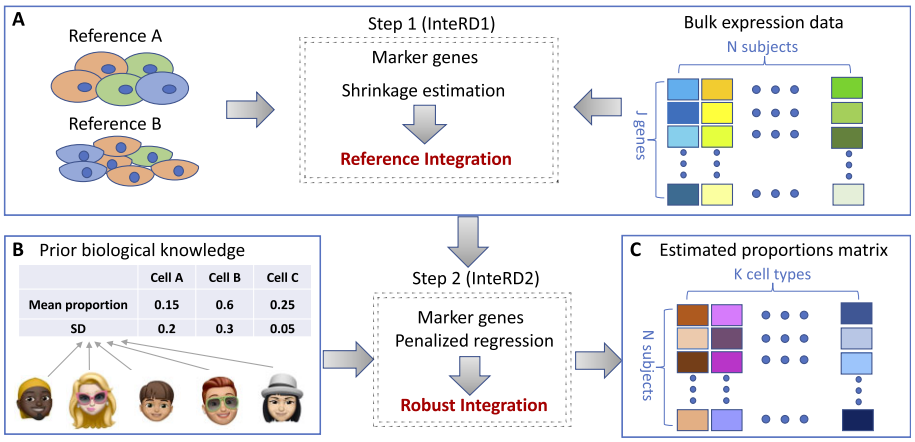

Fig. 1. A schematic workflow of InteRD. (**A**) InteRD first integrates proportion estimates based on multiple candidate reference sets. The input includes bulk expression data, the list of marker genes and reference sets, and the output is integrated RB estimates. (**B**) InteRD robustly aggregates the prior biological knowledge with integrated RB estimates by penalized regression and proposed evaluation criterion. The input includes RB estimates, the marker genes list and prior biological information extracted from scRNA-seq data or other independent studies; the output is the estimated proportions

reference data sets available for the target bulk tissue. Let $\tilde{\mathbf{P}}^{(r)}$ be the estimated cell-type proportions for reference dataset r. The RB integration framework aims to find the optimal sequence of weights ω_r that integrates all deconvolution results based on each reference in turn to approach the underlying truth $\mathbf{P}^T$, i.e.

$$\mathbf{P}^T \approx \omega_1 \tilde{\mathbf{P}}^{(1)} + \cdots + \omega_R \tilde{\mathbf{P}}^{(R)}, \tag{2}$$

where the sequence of non-negative weights is restricted to $\sum_{r=1}^{R} \omega_r = 1$ and leverages the priority among multiple reference sets to better approximate the underlying truth. However, as the true proportions $\mathbf{P}^T$ are unobserved, we cannot estimate these weights directly based on (2). Note that the gene expression matrix $\mathbf{Y}$ can be approximated by $\mathbf{P} \times \mathbf{M}$. Then, by applying the relationship in (2) and taking advantage of marker genes, we have

$$\mathbf{Y}^0 \approx \omega_1 \tilde{\mathbf{P}}^{(1)} \mathbf{M}^0 + \cdots + \omega_R \tilde{\mathbf{P}}^{(R)} \mathbf{M}^0, \tag{3}$$

where $\mathbf{Y}^0$ is a $N \times L$ expression matrix containing L marker genes in total for the cell types, and $\mathbf{M}^0$ is a $K \times L$ matrix containing unknown parameters for mean expression of marker genes. The above approximation is favored in the sense that the expression matrix $\mathbf{Y}^0$ as outcome in (3) is observed. Note that both ω_r and $\mathbf{M}^0$ are unknown. Thus, we introduced an iterative updating scheme to estimate both quantities. The overall estimation procedure starts with equal values of weights for each reference panel and can be summarized as follows.

Step 1.1: Given estimated weights $\hat{\boldsymbol{\omega}} = (\hat{\omega}_1, \ldots, \hat{\omega}_R)'$, InteRD updates the mean expression profile of gene j by applying augmented non-negative least squares (ANNLS) to the penalized regression from the loss $||\mathbf{Y}_j^0 - (\hat{\omega}_1 \tilde{\mathbf{P}}^{(1)} + \cdots + \hat{\omega}_R \tilde{\mathbf{P}}^{(R)}) \mathbf{M}_j^0||_2^2 + \lambda_1 ||\mathbf{S}_j' \mathbf{M}_j^0||_2^2$, where $\mathbf{M}_j^0 = (\mathbf{M}_{1j}^0, \ldots, \mathbf{M}_{Kj}^0)'$ is the vector for gene j containing the mean expression profile in each cell type; and $\mathbf{S}_j$ is a $K \times 1$ indicator vector with element i equal to zero if the gene j is regarded as a marker for cell type i and equal to one otherwise.

Step 1.2: Given mean expression profiles from marker genes, denoted by $\hat{\mathbf{M}}^0$, InteRD updates weights $\boldsymbol{\omega}$ by applying NNLS to the loss $||\mathbf{Y}^0 - \tilde{\mathbf{P}}^{(1)} \hat{\mathbf{M}}^0 \omega_1 - \ldots - \tilde{\mathbf{P}}^{(R)} \hat{\mathbf{M}}^0 \omega_R||_2^2$.

The penalty term and tuning parameter λ_1 in Step 1.1 shrink the cell-type-specific expression of gene j if this gene is not the marker for those cell types. As both $L2$ loss and a penalty are used, this shrinkage estimation can be easily implemented by ANNLS (Supplementary Section SC). The final estimates $\hat{\boldsymbol{\omega}}$ are determined by repeating the above two steps until there is little change in weight estimates. The advantage of InteRD for reference integration is that the whole process only uses the bulk expression and estimated cell-type proportions based on each reference, thus implicitly addressing the problem of batch-effect confounding from each reference set. The use of marker genes in (3) further facilitates reference ensemble by avoiding over-fitting and non-identifiability issue in this non-negative matrix factorization problem (Eggert and Korner, 2004). Although SCDC-ENSEMBLE started from the same idea in (2), it was implemented in a totally different manner, using distinct GEPs from each scRNA-seq data set to represent the underlying GEPs $\mathbf{M}$ in (3). Therefore, SCDC-ENSEMBLE still confounded the estimates of weights with expression profiles from multiple scRNA-seq datasets, as was also observed in our numerical studies. In addition, as SCDC-ENSEMBLE used GEPs from different reference panels, only the genes appearing in all reference sets were adopted, resulting in potential loss of information.

3.3 Step 2: Integration of RB results with other information

In addition to single-cell reference information, we considered two types of external information: marker genes identified by existing literature or scRNA-seq data; and the so-called population-level mean proportion for each cell type, a prior biological belief regarding the cell-type composition in a studied population. In practice, this information can be obtained from recognized biological findings or calculated from expression data from other studies, species or data types (Li *et al.*, 2020). Note that in the latter case we expect that the mean cell-type proportions as priors will be similar to those from the target bulk tissue. Now, let us denote the subject-level estimated cell-type proportion matrix via RB deconvolution by $\tilde{\mathbf{P}}$ and the vector containing population-level mean proportions by $\overline{\mathbf{P}}$. Both are assumed to be known. Then, the aggregating estimation procedure starts with equal proportions for each cell type and can be summarized as follows.

Step 2.1: Given estimated cell-type proportions, InteRD updates the mean expression profile of gene j as $\hat{M}_{kj}^0 = \sum_{i=1}^{N} Y_{ij}^0 / \sum_{i=1}^{N} \hat{P}_{ik}$ if $j \in \mathbf{W}_k$, and 0 otherwise. The set $\mathbf{W}_k$ is a collection of marker genes for cell type k.

Step 2.2: Given mean expression profiles for marker genes $\hat{\mathbf{M}}^0$ and grant variance $\hat{\sigma}^2$, InteRD updates cell-type proportions by minimizing the loss

$$(1/\hat{\sigma}^2)||\mathbf{Y}^0 - \mathbf{P} \times \hat{\mathbf{M}}^0||_2^2 + ||(\mathbf{P} - \overline{\mathbf{P}}')\boldsymbol{\Omega}||_2^2 + \lambda_2 ||\mathbf{P} - \tilde{\mathbf{P}}||_2^2, \tag{4}$$

where $\boldsymbol{\Omega}$ is a $K \times K$ diagonal matrix, of which the kth diagonal element is the inverse of the standard deviation of the kth cell-type proportion. This prior information can be also obtained from expression data from other studies. The overall variance σ^2 is

estimated by $||\mathbf{Y}^0 - \hat{\mathbf{P}} \times \hat{\mathbf{M}}^0||_2^2/|\mathbf{Y}^0|$ and is updated at each iteration, where $|\mathbf{Y}^0|$ denotes the cardinality of $\mathbf{Y}^0$. ANNLS is used to solve the penalized regression in (4). The final estimates $\hat{\mathbf{P}}$ are determined by repeating the above two steps until convergence.

Note that without the two penalty terms in (4), the above algorithm is reduced to a reference-free approach, whereas without the second penalty it is reduced to TOAST/+P (Li *et al.*, 2020), the partial reference-free approach. Note that TOAST/+P adopted a Bayesian idea by setting underlying cell-type proportions to follow a normal distribution as priors with hyperparameters: population-level mean proportions and standard deviations. The validity of TOAST/+P is based on the assumption that these hyperparameters calculated from external studies are accurate; nevertheless, this is hard to achieve and validate in practice, owing to differences in tissue preparation methods and heterogeneity among studied subjects. Thus, adding the second penalty in (4) further enriches the deconvolution system incorporating RB estimates and also provides double protection against incorrect information in the estimation. The strength of information borrowing and protection is controlled by the tuning parameters λ_2, which should be tuned based on the data.

3.4 Tuning parameters

To effectively solve the system described in Section 3.2 and integrate different information into the deconvolution in a manner that is robust against any incorrect prior information as described in Section 3.3, the selection of tuning parameters λ_1 and λ_2 is critical. We developed the following lack-of-fit-based metric for both optimization problems. Briefly, for any estimation matrix $\hat{\mathbf{P}}$ based on given tuning parameters, we conducted NNLS regression to fit $\mathbf{Y}^c$, the whole expression data excluding marker genes as validation data to alleviate the issue of overfitting. By solving this regression problem, we obtained estimates of the mean expression profiles, denoted by $\hat{\boldsymbol{\mu}}^c$. Thereafter, the evaluation was calculated by the L_1 norm between responses $\mathbf{Y}^c$ and imputed outcomes $\hat{\mathbf{P}} \times \hat{\boldsymbol{\mu}}^c$. The tuning parameters were determined based on the smallest value of the proposed criterion via a grid search. The underlying goal when selecting tuning parameters is to capture proper information yielding a good estimation of cell-type proportions that best fits the target bulk data. Thus, InteRD enables a data-driven mechanism to robustly aggregate prior biological knowledge into deconvolution. The pseudo-codes for the algorithms and further discussion about the selection of tuning parameters are provided in the Supplementary Material.

4 Simulation studies

4.1 Pseudo-bulk data

In this section, we describe the extensive simulations that were conducted to assess the utility and robust properties of the InteRD algorithm. To mimic the real-world situation, we simulated pseudo-bulk data from pancreas scRNA-seq data generated from six healthy donors by Segerstolpe *et al.* (2016). Specifically, a mixed scRNA-seq reference was first constructed by pooling cells across all donors, which ensured enough variability among simulated samples. Then, for each cell type (alpha, beta, delta and gamma), we sub-sampled cells without replacement from the mixed reference, with the sampling rate generated by a uniform distribution $U(0.3, 0.7)$. After re-sampling cells from all cell types, the pseudo-bulk sample was created by aggregating sampled cells, where cell-type proportions were calculated based on cell abundance in the re-sampled data. Finally, we repeated the re-sampling scheme $N = 25$, 40, 80 times to generate multiple pseudo-bulk samples. More evaluations based on other simulated or existing bulk data could be found in the Supplementary Material.

4.2 Single-cell references, marker genes and other external information

To explicitly demonstrate the utility of InteRD, we separately show results based on one-step InteRD (named InteRD1 and described in Section 3.2) and two-step InteRD (named InteRD2 and described in Section 3.3). InteRD1 integrates estimated cell-type proportions derived from two different scRNA-seq references, one from Baron *et al.* (2016) and the other one from Xin *et al.* (2016) (12 healthy donors). We ran SCDC to obtain estimated proportions for each reference; it showed better performance than other single RB deconvolution methods (Dong *et al.*, 2021). More evaluations about comparing InteRD with other deconvolution algorithms based on single reference with/without batch correction can be found in Supplementary Section SE. On the other hand, InteRD2 aggregates the results from InteRD1 with population-level mean proportions $\overline{\mathbf{P}}$ and a reference-free approach. Note that both InteRD1 and InteRD2 require a list of marker genes. We separately ran InteRD based on each set of candidate marker genes selected from Baron *et al.* (2016) and Xin *et al.* (2016), respectively, by statistically testing genes over-expressed in one cell type compared with the remaining types based on the package *Seurat*. Note that neither the selected scRNA-seq datasets nor the marker genes were paired with the pseudo-bulk data; this represented a challenging but more realistic scenario, as library preparation protocols may vary across studies, and deconvolution of the target bulk tissue and marker gene selection in practice is often performed based on external reference data. To assess the robustness of InteRD2 to prior biological information, we considered two candidates of population-level mean proportions and corresponding standard deviations: one was calculated from the source of the scRNA-seq data in Segerstolpe *et al.* (2016) study, which mimicked the case where the prior information was correct; the other was calculated from the estimated cell-type proportions based on Xin *et al.* (2016) as a reference. The latter deviated from the underlying truth and mimicked the case where the prior information was inaccurate. In practice, population-level mean proportions and their standard deviations over individuals could substantially vary between underlying bulk data and scRNA-seq data from independent studies owing to differences in tissue preparation methods, heterogeneity among studied subjects and other uncontrolled experimental factors. Thus, it is always risky to impose a pre-determined structure into bulk tissue data.

4.3 Evaluations

We evaluated the robustness of InteRD scheme in the presence of multiple single-cell reference datasets (potentially) inaccurate population-level mean proportions $\overline{\mathbf{P}}$ and/or marker genes. We conducted 100 Monte Carlo runs with various sample sizes and different numbers of marker genes per cell type. To make consistent assessments, we compared InteRD1 and InteRD2 with the reference-free approach (Ref-free), the single RB SCDC, SCDC-ENSEMBLE and TOAST/+P. Note that the existing methods only used partial information and relied on the assumption that the information used should be accurate. For example, the reference-free approach only involved marker genes in the deconvolution, whereas SCDC and SCDC-ENSEMBLE only incorporated information from single-cell references. TOAST/+P, on the other hand, used marker genes and prior composition knowledge ($\overline{\mathbf{P}}$) to guide cell-type proportion estimation. We used mean absolute deviation (MAD) averaged over samples and Monte Carlo runs and Kendall rank correlation coefficient (Kendall) averaged over Monte Carlo runs to assess the similarity between recovered cell-type proportions and the underlying truth. The lower MAD the better, whereas the higher Kendall the better. A method that simultaneously achieves low MAD and high Kendall is much more preferred.

Cell-type-specific results are shown in Figure 2 (Supplementary Fig. S13 for more details) with a sample size of 40, and 20 markers per cell type selected from Xin *et al.* (2016). In general, InteRD1 and InteRD2 outperformed other algorithms in terms of smaller MAD and higher Kendall over 100 Monte Carlo runs. InteRD1 successfully integrated two estimated cell-type proportions, of which the resulting estimates were close to the better estimates based on the scRNA-seq reference from Baron *et al.* (2016). This signifies the ability of InteRD1 to distinguish multiple reference datasets and to identify the best one. However, SCDC-ENSEMBLE failed to capture the proper reference (Baron *et al.*, 2016) and led to substantial

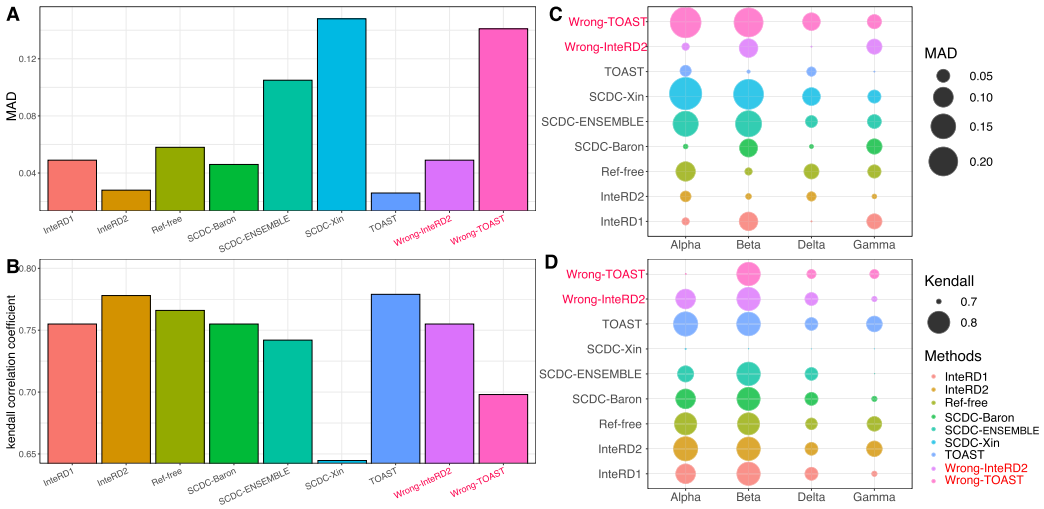

Fig. 2. Comparison of cell-type proportion estimation for pseudo bulk data ($n = 40$) generated from Segerstolpe *et al.* (2016) based on mean absolute difference (MAD, the smaller the better) averaged over samples and 100 Monte Carlo runs and Kendall rank correlation coefficient (Kendall, the higher the better) averaged over 100 Monte Carlo runs. Twenty markers per cell type were selected from Xin *et al.* (2016). (**A** and **B**) Overall evaluations based on averaged MAD and Kendall over cell types, respectively; (C and D) cell-type-specific evaluations with circle sizes representing MAD values and Kendall values, respectively. InteRD2 (wrong-pbar) and TOAST (wrong-pbar) in red represent the estimates given incorrect biology information as prior (A color version of this figure appears in the online version of this article.)

MAD and lower Kendall. Better performance of InteRD1 over SCDC-ENSEMBLE was also observed under another setup, where pseudo-bulk data were generated by sampling scRNA-seq data from Baron *et al.* (2016) (Supplementary Fig. S1). Among methods not using scRNA-seq data as a reference, the reference-free approach led to higher MAD compared with InteRD1 and was observed to be more sensitive to sample size (Supplementary Tables S1 and S2, Supplementary Figs S12–S17). When the population-level mean proportion and standard deviations were calculated from the true source, TOAST/+P led to satisfactory estimates, whereas it lost its power and resulted in more biased estimates when the prior knowledge deviated from the underlying pseudo-bulk data. On the other hand, InteRD2 further improved on the performance of InteRD1 by aggregating information from marker genes and accurate population-level mean proportions. It had the lowest MAD and the highest Kendall among all methods. Even the prior knowledge about population-level proportions deviated from the underlying source; InteRD2 activated its navigation by searching for the best λ in (4) and finally gave most credit to InteRD1 estimates. The same patterns were also observed when marker genes were selected from Baron *et al.* (2016) (Supplementary Fig. S2). We also evaluated the impact of the sample size and number of marker genes on deconvolution (Supplementary Tables S1 and S2). The InteRD scheme proved to be robust and always led to satisfactory results under various settings. Superiority of InteRD scheme over existing methods was also detected under the pseudo-bulk data simulated from dorsolateral prefrontal cortex tissue with seven cell types (Supplementary Fig. S3) and under the brain cortical bulk data with cell-type proportions calculated by an orthogonal method (immunohistochemistry) (Supplementary Figs S10 and S11).

5 Real-data applications

5.1 Human pancreatic islets

To further evaluate the utility of the InteRD scheme, we used bulk RNA-seq data from islets of Langerhans, groups of endocrine cells within the pancreas, which are essential for blood glucose homeostasis. We conducted deconvolution analysis for the bulk RNA-seq data from (Fadista *et al.*, 2014), which included 77 subjects, of whom 51 were considered to be healthy (hbA1C level ≤ 6) and 26 had T2D (hbA1C level > 6). Owing to the availability of multiple reference datasets, we adopted InteRD1 (defined in Section 4.2) to integrate multiple reference sets, and compared the performance of InteRD1 with a state-of-the-art data integration algorithm, SCDC-ENSEMBLE. Two single-cell reference datasets, Baron *et al.* (2016) and Segerstolpe *et al.* (2016) were used as reference candidates. For marker gene selection, we adopted a similar statistical test strategy to that described in Section 4 and also added genes already recognized as well-known markers in the literature (Segerstolpe *et al.*, 2016). Statistical tests based on the package *Seurat* were used to select 60 genes over-expressed in one cell type from each scRNA-seq reference dataset, respectively. The final set was then determined by taking the intersection of candidate genes from these two datasets. To account for sample heterogeneity, we applied InteRD1 and SCDC-ENSEMBLE separately for healthy and T2D tissues. The results are summarized in Figure 3A. InteRD1 consistently gave all weights to the reference from Segerstolpe *et al.* (2016), regardless of healthy or T2D grouping. Weights selected by SCDC-ENSEMBLE were nevertheless mixed and inconsistent across the two groups, with increased weights assigned to Baron *et al.* (2016) in T2D tissues (Supplementary Fig. S6). To further benchmark the performance, we focused on beta cells, which have been shown to be gradually lost during T2D progression (Dong *et al.*, 2021; Wang *et al.*, 2019). Note that InteRD1 implied 50% beta cell proportions in healthy samples, which agreed well with previous findings that beta cell proportions are expected to be around 50% in healthy adult human pancreatic islets. On the other hand, the reference-free approach recovered small beta cell proportions in both healthy and T2D tissues. To integrate additional information, we used InteRD2 (defined in Section 4.2) and TOAST/+P. The population-level mean proportions $\overline{\mathbf{P}}$ and corresponding standard deviations across subjects were calculated based on scRNA-seq data from Baron *et al.* (2016) and Segerstolpe *et al.* (2016). As shown in Figure 3A, the proportion estimates from InteRD2 were the same as those from InteRD1 in healthy tissues and slightly different in T2D tissues. These results implied that InteRD2 treated the prior biological information as inappropriate for the system, especially in healthy tissues. However, TOAST/+P led to totally different results compared with InteRD2, with lower beta cell proportions detected in healthy tissues and elevated beta cell proportions in T2D tissues. It could be the case that TOAST/+P failed to judge the reliability of the provided biological information. This limitation was also observed in the numerical evaluations presented in Section 4.

To replicate previous findings that beta cells are gradually lost during progression of T2D, we constructed a multivariate regression model using estimated beta cell proportions as outcomes and other covariates, including age, gender, body mass index and HbA1c, as

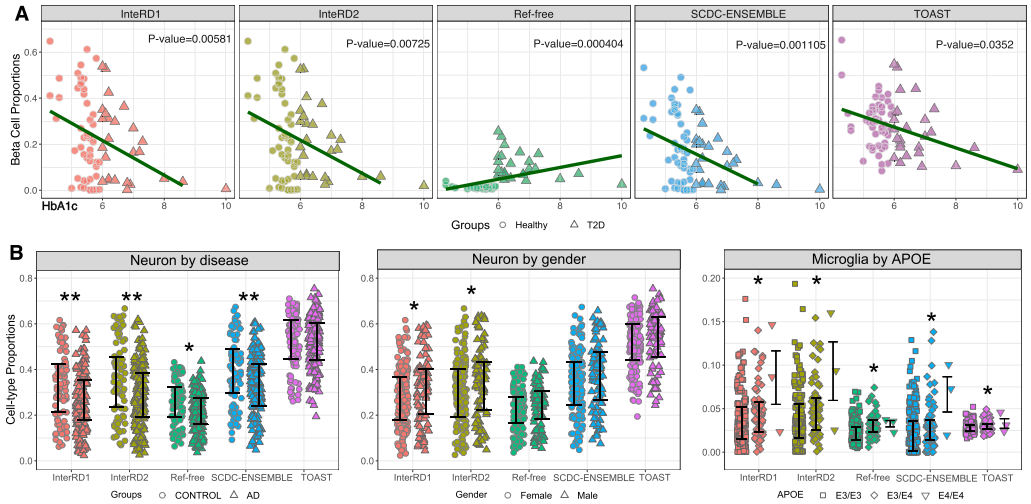

Fig. 3. (**A**) Pancreatic islet cell-type composition in healthy and T2D human samples. Scatter plots for HbA1c versus beta cell-type proportions recovered by five color-coded by deconvolution methods. *P*-values are provided to evaluate statistical association between beta cell proportions and HbA1c. (**B**) Dorsolateral prefrontal cortex cell-type composition in AD and control groups. The Jitter plot to the left includes neuron cell-type proportions for ROSMAP, color-coded by deconvolution methods and shape-coded by disease conditions. The Jitter plot in the middle includes neuron cell-type proportions shaped-coded by Gender. The Jitter plot to the right includes microglia cell-type proportions, shape-coded by the APOE genotype. In each plot, black bars indicate the range from the lower quartile to the upper quartile. **P*-value <0.05, ***P*-value <0.01

predictors. Note that HbA1c is an important biomarker for T2D, with a larger value indicating higher disease risk. As shown in Figure 3A, InteRD1, InteRD2 and SCDC-ENSEMBLE showed a strongly significant negative association between beta cell proportions and HbA1c with *P*-values <0.01, whereas TOAST/+P showed weak significance, and the reference-free approach failed to detect this negative association. Consistent patterns were also found by the InteRD estimates in the sub-cohort analysis by only considering T2D samples (Supplementary Fig. S7).

5.2 Human dorsolateral prefrontal cortex

We also conducted a deconvolution analysis for samples in the Rush Religious Orders Study and Memory and Aging Project (ROSMAP), which consists of postmortem brains from 155 donors with AD (cogdx= 4, braaksc $\geq$ 4, ceradsc $\leq$ 2) and 86 control donors (cogdx= 1, braaksc $\leq$ 3, ceradsc $\geq$ 3) (De Jager *et al.*, 2018). For external datasets, we used snRNA-seq data generated from middle frontal neocortex brain tissue (Nguyen *et al.*, 2020) and prefrontal cortex (Mathys *et al.*, 2019). Information about samples with neurofibrillary degeneration (A+T+), samples without neurofibrillary tangles but with beta-amyloid (A+T−), and samples without both neurofibrillary tangles and beta-amyloid (A−T−) was available in the former reference (Nguyen *et al.*, 2020), whereas the information about AD and control donors was available in the latter reference. These five subgroups in total may show different GEPs owing to the various disease conditions and thus were better treated as separate reference sets. Furthermore, to account for sample heterogeneity in the underlying bulk data, we applied all deconvolution methods separately for AD and control donors by integrating five snRNA-seq subsamples as five candidate reference sets. In a process similar to that described in Section 5.1, the marker genes were selected using well-known findings (Nguyen *et al.*, 2020) and statistical tests based on two reference sets. Interestingly, InteRD1 and SCDC-ENSEMBLE led to similar results, consistently selecting the snRNA-seq data with AD donors in Mathys *et al.* (2019) as the most proper reference set for both the AD group and control group in the bulk data (Supplementary Fig. S6). To obtain prior biological information, we calculated the population-level mean proportions and corresponding standard deviations for seven cell types based on snRNA-seq data from Nguyen *et al.* (2020) and Mathys *et al.* (2019). InteRD2 and TOAST/+P were used to aggregate this information.

To benchmark the performance of different cell-type deconvolution methods, we focused on neurons (combining excitatory and inhibitory neurons) and microglia, two cell types that are known to be associated with neurodegeneration and cognitive disorders. We compared InteRD2 with the reference-free approach, TOAST/+P, SCDC-ENSEMBLE and InteRD1. As shown in Figure 3B, the neuron proportions estimated by InteRD1, InteRD2, the reference-free approach and SCDC-ENSEMBLE showed statistically significant decreases in the AD group compared with the control group, based on multivariate regression after adjusting for age, gender, APOE, and education. TOAST/+P failed to detect decreased neuron proportions in the AD group. InteRD1 and InteRD2 also matched previous findings of statistically lower neuron proportions in female donors compared with male donors (Rabinowicz *et al.*, 2002). On the other hand, all methods successfully detected the negative effects of APOE (coded 0, 1 and 2 based on the frequency of the E4 risk allele) on microglia proportions after adjusting for disease condition, age, gender and education. Note that the patterns from InteRD1 and InteRD2 were rather consistent, with slightly higher neuron proportions and lower endothelial cell proportions recovered by InteRD2 (Supplementary Fig. S5). This difference was due to the integration of prior biological information in a robust manner. It is also worth mentioning that the estimates from InteRD2 were in concordance with findings in the literature (Wang *et al.*, 2020), and InteRD scheme is computationally much more efficient than SCDC-ENSEMBLE in the presence large number reference panels (Supplementary Table S4).

6 Discussion

With the explosion in the amounts of publicly available transcriptomics data arising from recent technological advances in bulk RNA-seq and scRNA-seq, it is becoming interestingly important to integrate multiple scRNA-seq reference sets and information from various sources to facilitate cell-type deconvolution. Also, owing to platform differences, sample heterogeneity and other inherent biases in sequencing data, caution is needed when integrating data generated from different sources. Failure to account for experimental noise and bias from disparate studies may worsen the performance of cell-type deconvolution algorithms. Current available deconvolution methods can be broadly classified as RB or reference-free approaches. This work unifies these two different approaches by

proposing a computational scheme to robustly and effectively combine prior biological information from existing sources.

The InteRD algorithm can be further improved in several ways. Instead of NNLS, some parametric approaches, such as Negative Binomial regression, can be considered in InteRD framework to better account for mean–variance relationship in count data. However, in the presence of multiple reference panels and prior biological information, it is computationally very challenging due to the facts of constrained cell-type proportions, non-negative gene expression values, and their linear relationship, which merits future work. Also, marker gene selection is an important step for InteRD. Currently, marker genes are selected from external scRNA-seq data based on statistical tests and prior biological findings reported in the literature. It would be desirable to develop a data-driven approach to select marker genes and integrate them into InteRD. One potential is to down weight those genes that could be 'aberrant' markers by means of the technique in Wilson *et al.* (2020). The absence of certain rare cell types from external sources is another practical issue that one may encounter in real applications. How to make use of datasets where some cell types are missing is worth further investigation. In summary, we have developed a unified framework for cell-type deconvolution that can efficiently integrate multiple reference sets and is robust to potential inaccuracies in prior biological information. We believe that the proposed computational scheme is a more flexible and feasible alternative to existing deconvolution methods and will facilitate biological discoveries in many applications.

Funding

This work was supported by the following grants: National Institute of General Medical Sciences [R01GM125301 to M.L.]; National Heart, Lung, and Blood Institute [R01HL113147 and R01HL150359 to M.L.]; National Eye Institute [R01EY030192, R01EY031209 and R21EY031877 to M.L.]; and National Institute on Aging [U24 AG041689 and U54 AG052427 to L.-S.W.]. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflict of Interest: none declared.

References

Baron,M. *et al.* (2016) A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Syst.*, **3**, 346–360.

Chen,X. *et al.* (2018) From tissues to cell types and back: single-cell gene expression analysis of tissue architecture. *Annu. Rev. Biomed. Data Sci.*, **1**, 29–51.

De Jager,P.L. *et al.* (2018) A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci. Data*, **5**, 1–13.

Dong,M. *et al.* (2021) SCDC: bulk gene expression deconvolution by multiple single-cell rna sequencing references. *Brief. Bioinform.*, **22**, 416–427.

Eggert,J. and Korner,E. (2004) Sparse coding and NMF. In: *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541), Budapest, Hungary*, Vol. 4. IEEE, pp. 2529–2533.

Fadista,J. *et al.* (2014) Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc. Natl. Acad. Sci. USA*, **111**, 13924–13929.

Houseman,E.A. *et al.* (2014) Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, **30**, 1431–1439.

Hrdlickova,R. *et al.* (2017) RNA-seq methods for transcriptome analysis. *Wiley Interdiscip. Rev. RNA*, **8**, e1364.

Kuhn,A. *et al.* (2011) Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nat. Methods*, **8**, 945–947.

Li,Z. *et al.* (2020) Robust partial reference-free cell composition estimation from tissue expression. *Bioinformatics*, **36**, 3431–3438.

Mathys,H. *et al.* (2019) Single-cell transcriptomic analysis of Alzheimer's disease. *Nature*, **570**, 332–337.

Menden,K. *et al.* (2020) Deep learning-based cell composition analysis from tissue expression profiles. *Sci. Adv.*, **6**, eaba2619.

Newman,A.M. *et al.* (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, **12**, 453–457.

Newman,A.M. *et al.* (2019) Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.*, **37**, 773–782.

Nguyen,A.T. *et al.* (2020) APOE and TREM2 regulate amyloid-responsive microglia in Alzheimer's disease. *Acta Neuropathol.*, **140**, 477–493.

Rabinowicz,T. *et al.* (2002) Structure of the cerebral cortex in men and women. *J. Neuropathol. Exp. Neurol.*, **61**, 46–57.

Repsilber,D. *et al.* (2010) Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach. *BMC Bioinformatics*, **11**, 1–15.

Segerstolpe,Å. *et al.* (2016) Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.*, **24**, 593–607.

Wang,X. *et al.* (2019) Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.*, **10**, 1–9.

Wang,X. *et al.* (2020) Deciphering cellular transcriptional alterations in Alzheimer's disease brains. *Mol. Neurodegeneration*, **15**, 1–15.

Wilson,D.R. *et al.* (2020) Iced-t provides accurate estimates of immune cell abundance in tumor samples by allowing for aberrant gene expression patterns. *J. Am. Stat. Assoc.*, **115**, 1055–1065.

Xin,Y. *et al.* (2016) RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab.*, **24**, 608–615.