EasyChair Preprint

№ 11132

Imply: Improving Cell-Type Deconvolution Accuracy Using Personalized Reference Profiles

Guanqun Meng, Yue Pan, Wen Tang, Lijun Zhang, Ying Cui, Fredrick R. Schumacher, Ming Wang, Rui Wang, Sijia He, Jeffrey Krischer, Qian Li and Hao Feng

October 23, 2023

imply: improving cell-type deconvolution accuracy using personalized reference profiles

Guanqun Meng[1], Yue Pan[2], Wen Tang[1], Lijun Zhang[1], Ying Cui[3], Fredrick R. Schumacher[1], Ming Wang[1], Rui Wang[4], Sijia He[5], Jeffrey Krischer[6], Qian Li[2,*], and Hao Feng[1,*]

[1] Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, 44106, USA
[2] Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, 38105, USA
[3] Department of Biomedical Data Science, Stanford University, Stanford, CA, 94305, USA
[4] Department of Surgery, Division of Surgical Oncology, University Hospitals Cleveland Medical Center, Cleveland, OH, 44106, USA
[5] Department of Biostatistics, University of Michigan, Ann Arbor, MI, 48109, USA
[6] Health Informatics Institute, University of South Florida, Tampa, FL, 38105, USA

Abstract. Real-world clinical samples are often admixtures of signal mosaics from multiple pure cell types. Using computational tools, bulk transcriptomics can be deconvoluted to solve for the abundance of constituent cell types. However, existing deconvolution methods are conditioned on the assumption that the whole study population is served by a single reference panel, which ignores person-to-person heterogeneity. Here we present imply, a novel algorithm to deconvolute cell type proportions using personalized reference panels. imply can borrow information across repeatedly measured samples for each subject and obtain precise cell type proportion estimations. Simulation studies demonstrate reduced bias in cell type abundance estimation compared with existing methods. Real data analyses on large longitudinal consortia show more realistic deconvolution results that align with biological facts. Our results suggest that disparities in cell type proportions are associated with several disease phenotypes in type 1 diabetes and Parkinson's disease. Our proposed tool imply is available through the R/Bioconductor package ISLET at https://bioconductor.org/packages/ISLET/.

Keywords: Deconvolution · Bulk RNA-seq · Personalized reference · Admixed samples · Cell-type-specific.

1 Background

Tissues are complex samples composed of different cell types, and real bulk transcriptomic data are often weighted sums of multiple signals over several different cell types [18]. In large-scale and population-level clinical studies, like Parkinson's Disease Biomarkers Program (PDBP) and The Cancer Genome Atlas (TCGA), transcriptomic samples are often collected from complex tissues. For admixed tissue samples, differentially expressed transcriptional profiles from different phenotypical groups can be caused by either cell-type composition disparities or underlying cell-type-specific gene expression heterogeneity. Studies have shown that cell type proportions are confounders with other phenotypical covariates like age, sex, or clinical outcomes for bulk transcriptomic data analysis [5, 6]. As a result, ignoring cell-type-specific compositions in gene expression analysis would cause inflated false positive rates of identifying relevant genetic features. An accurate cell type proportion deconvolution is thus vital, especially for cell types with low abundance and weak biological signals where the real biological differences could be shadowed by technical noises [28, 6, 39].

Recently, several statistical methods have been proposed to deconvolute cell type abundance from bulk transcriptome data. These methods utilize the statistical framework of linear least squares regression [52, 58, 12], quadratic programming [24], support vector regression [43, 11], and non-negative matrix factorization [20, 46]. These methods share the same goal of quantifying the unknown abundances of various cell types and can be broadly summarized into two categories: Reference-Based (RB) and Reference-Free (RF). The RB deconvolution relies on a cell-type-specific gene expression signature reference panel composed of the preselected features known to differentiate cell types, while the RF deconvolution estimates cell type proportions in the absence of a reference panel. In general, RB approaches have better performance compared with RF approaches [5]; however, the accuracy of cell type abundance inference is dependent on the quality of signature matrices [5]. RF deconvolution, in contrast, offers flexibility where reference panels are hard to obtain.

RB deconvolutions require a reference panel as the input. CIBERSORT [43], which is a state-of-art RB deconvolution approach [5, 6], provides a verified signature panel LM22. It is specifically for leukocyte deconvolution and includes 547 marker genes for 22 hematopoietic cell types. xCell [4] combines the gene set enrichment with deconvolution techniques and introduces curated gene signatures representing 64 distinct cell types. However, it is a very strong assumption to use a single reference panel across the whole population and ignore person-to-person heterogeneity. It also deviates from the biological fact that the gene expression profile could vary, even for one purified cell type, depending on environmental influences, age, sex, subject's health status, and treatment paradigms [27, 14, 17, 26, 1, 22, 9, 51, 40]. Mismatched reference signatures can impact deconvolution results [50, 21]. The problem is even exacerbated when handling longitudinal data, when intra-subject samples share information and inter-subject heterogeneities are relatively strong. Recent research shows that models incorporating personalized effects can accurately retrieve cell type reference panels on the individual-basis [16]. However, to date, no method is available to take advantage of personalized references panel to precisely deconvolute cell type proportions, especially when longitudinal samples are available.

Here we develop a new deconvolution algorithm ***imply*** (***imp****roving cell-****ty****pe deconvolution using personalized reference*) as depicted in **Figure 1**. ***imply*** can utilize personalized reference panels to precisely deconvolute cell type proportions using longitudinal data. It borrows information across the repeatedly measured transcriptome samples within each subject, to recover personalized reference panels. The personalized references are further adopted to improve cell type deconvolution. The rationale of our approach is straightforward: the personalized reference panel is more accurate compared with the population-level signature and using a personalized reference panel can consequently lead to a more precise cell-type deconvolution.

We conducted extensive *in silico* simulations and real data analyses to test the performance of ***imply***. Simulation results showed that ***imply*** reduced bias in deconvolution and increased the correlation between the estimated and ground-truth cell type abundance. Real data analyses on two large longitudinal consortia, The Environmental Determinants of Diabetes in the Young (TEDDY) and Parkinson's Disease Biomarkers Program (PDBP), showed more realistic results that align with low-throughput experiments. The results suggested that disparities in cell type proportions of certain cell types are associated with type 1 diabetes and Parkinson's disease. Our method ***imply*** has been implemented and integrated into the Bioconductor package *ISLET* and is available at https://bioconductor.org/packages/ISLET/.

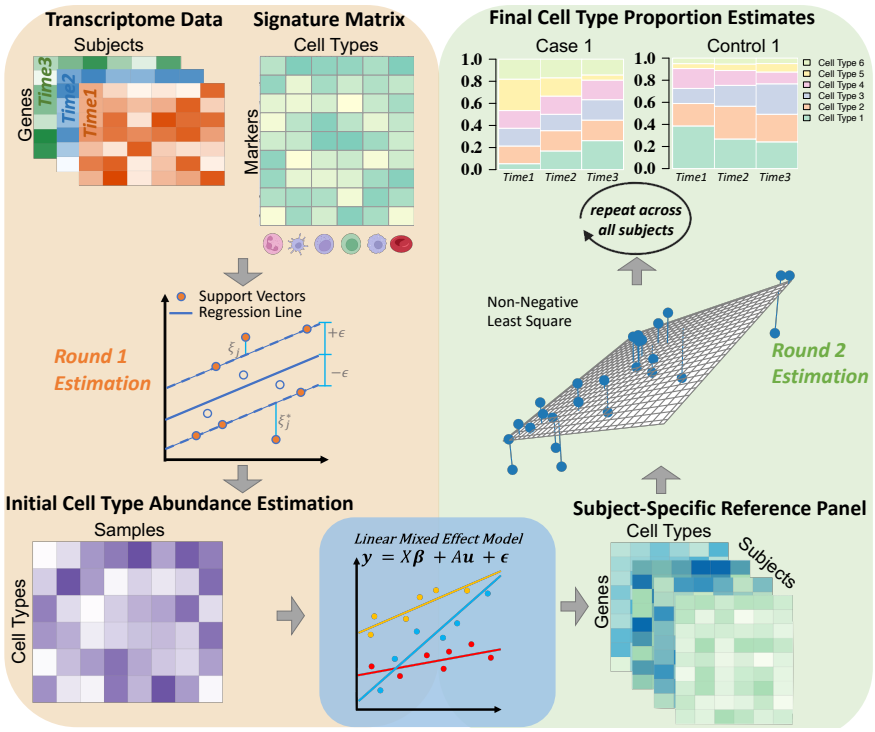

Fig. 1. Overview of ***imply***'s personalized deconvolution. The **top-left** shows two inputs: repetitively measured transcriptome data and a signature matrix containing cell-type-specific marker genes. *Stage I*, depicted in the **middle left**, adopts support vector regression to derive a preliminary cell type abundance, as shown in the **bottom-left**. Next, for *Stage II*, as shown in the **bottom-center**, linear mixed effect models are utilized to reconstruct personalized references, which are shown in the **bottom-right**. In *Stage III*, as illustrated in the **middle-right**, by employing *non-negative least square* and personalized references generated from the previous step, repeatedly across all subjects, ***imply*** enables personalized deconvolution to produce cell type proportion estimates, shown on the **top-right**.

2 Methods

Overview of *imply* To outline briefly, the primary objective of ***imply*** is to improve the accuracy of cell abundance estimations through the integration of subject- and cell-type specific reference panels, termed personalized references. The algorithm is structured into three stages. In *Stage I*, the initial cell proportion estimates will be obtained. The core component of ***imply*** lies in *Stage II*, where a personalized reference panel is retrived for each subject. These personalized references will replace the population-level signature matrix, facilitating a personalized deconvolution process repeatedly across all subjects in *Stage III*.

Notation introduction Let $G(g = 1, 2, 3, \dots G)$ denotes the total number of features (e.g., genes), and $N(n = 1, 2, 3, \dots N)$ as the total number of subjects. For each subject n, there are $t_n(i = 1, 2, \dots t_n)$ repeated measurements. The total number of samples across N subjects is thus $T = \sum_{n=1}^{N} t_n$. The bulk transcriptome dataset can be represented as a matrix $\boldsymbol{Y}$, of dimension $G \times T$. We denote $K(k = 1, 2, \dots K)$ as the total number of purified cell types. Initially, we would have a population-level signature matrix $\boldsymbol{E}$ of dimension $J \times K(J < G)$, where J indicates the total number of discriminative signature genes for the first-round cell type deconvolution in *Stage I*. This signature matrix can be derived from pure cell line data or aggregated from annotated single-cell RNA-seq (scRNA-seq) data [55, 25, 15].

2.1 Stage I: Initial cell type proportion estimation

With the observed admixed data $\boldsymbol{Y}$ and the initial reference panel $\boldsymbol{E}$, as illustrated in the top-left of **Figure 1**, the first-round RB coarse deconvolution is conducted using a ν-Support Vector Regression algorithm (ν-SVR) [48] based on a linearity assumption [29, 49]. This strategy was already proven to be a successful choice in deconvolution algorithms such as CIBERSORT [43]. Support vectors are regulated by ϵ-tubes integrated

into the objective function (specified below). The deconvolution is modeled by a regression problem: $\boldsymbol{y}_{\cdot ni} = f(\boldsymbol{\theta}_{E,ni\cdot}) = \boldsymbol{E}\boldsymbol{\theta}_{E,ni\cdot} + \boldsymbol{b}$, where $\boldsymbol{b} \in R^J$ capture random bias, and we can minimize the following objective function [43]:

$$\text{argmin}\, \frac{1}{2}\|\boldsymbol{\theta}_{E,ni\cdot}\|^2 + C\sum_{j=1}^{J}(\xi_j + \xi_j^*), \quad \xi_j, \xi_j^* > 0$$

The solved $\hat{\boldsymbol{\theta}}_{E,ni\cdot} = [\hat{\theta}_{E,ni1}, \hat{\theta}_{E,ni2}, \dots \hat{\theta}_{E,niK}]'$ is the first-round sample-specific cell type abundance estimation. The constraints of the objective function and parameters of ϵ, C, ξ_j, and ξ_j^* are detailed in the supplementary material section 1.1. Then negative coefficients ($\hat{\boldsymbol{\theta}}_{E,ni\cdot}$) are set to 0, and the remaining coefficients are normalized to sum-to-one, which is the general practice in proportion deconvolution [43]. Repeating this process for all samples, we obtain the deconvoluted cell composition matrix $\hat{\boldsymbol{\Theta}}_E$ with dimension $T \times K$. It's worth to note that this first-step deconvolution of cell type proportions provides a valid foundation for downstream steps.

$$\hat{\boldsymbol{\Theta}}_E = \begin{bmatrix} \hat{\theta}_{E,111} & \hat{\theta}_{E,112} & \dots & \hat{\theta}_{E,11K} \\ \hat{\theta}_{E,121} & \hat{\theta}_{E,122} & \dots & \hat{\theta}_{E,12K} \\ \vdots & & \ddots & \\ \hat{\theta}_{E,1t_11} & \hat{\theta}_{E,1t_12} & \dots & \hat{\theta}_{E,1t_1K} \\ \vdots & & \ddots & \\ \hat{\theta}_{E,N11} & \hat{\theta}_{E,N12} & \dots & \hat{\theta}_{E,N1K} \\ \hat{\theta}_{E,N21} & \hat{\theta}_{E,N22} & \dots & \hat{\theta}_{E,N2K} \\ \vdots & & \ddots & \\ \hat{\theta}_{E,Nt_N1} & \hat{\theta}_{E,Nt_N2} & \dots & \hat{\theta}_{E,Nt_NK} \end{bmatrix}$$

2.2 Stage II: Personalized reference panel recovery

The second step aims to retrieve a subject- and cell-type-specific reference panel. Using the cell-type-specific and sample-specific proportions $\hat{\theta}_{E,nik}$ from *Stage I*, we can set up the linear mixed-effect regression for each gene g: $\boldsymbol{y}_{g\cdot\cdot} = \boldsymbol{X}\boldsymbol{\beta}_g + \boldsymbol{A}\boldsymbol{u}_g + \boldsymbol{\epsilon}_g$, where $\boldsymbol{\epsilon}_g \sim N(\boldsymbol{0}, \sigma_0^2\boldsymbol{I})$ are the residuals and $\boldsymbol{y}_{g\cdot\cdot}$ is the vector of observed expression data for a specific gene g for all T samples. $\boldsymbol{X}$ and $\boldsymbol{A}$ are the design matrices (each with dimension $T \times 2K$ and $T \times NK$) for the fixed-effect $\boldsymbol{\beta}_g$ and the random-effect $\boldsymbol{u}_g$ (details of $\boldsymbol{X}$ and $\boldsymbol{A}$ are specified in supplementary material section 1.2). The initial cell type abundance information is further reorganized into vectors $\boldsymbol{a}_{nk} = (\hat{\theta}_{E,n1k}, \hat{\theta}_{E,n2k}, \dots \hat{\theta}_{E,nt_nk})'$. The fixed-effect $\boldsymbol{\beta}_g = (m_1, m_2, \dots m_K, \beta_1, \beta_2, \dots \beta_K)'$ has two components: $(m_1, m_2, \dots m_K)$ are the baseline average cell-type-specific gene expression in the control group, and $(\beta_1, \beta_2, \dots \beta_K)$ are the '*difference*' between the case group and the control group at cell type level. Note that our modeling allows for the incorporation of subject-level covariates such as disease status, $z_n = 1/0$ for disease or normal. The random-effect $\boldsymbol{u}_g = (u_{11}, u_{21}, \dots u_{N1}, u_{12}, u_{22}, \dots u_{N2}, \dots, u_{1K}, u_{2K}, \dots u_{NK})'$ captures the subject-level and cell-type-specific gene expression deviation from the group-level average. $\hat{\boldsymbol{\beta}}_g$ and $\hat{\boldsymbol{u}}_g$ can be solved by penalized least square algorithm with restricted maximum likelihood [8]. The subject- and cell-type-specific reference panel (denoted as $\boldsymbol{R}_n$ with dimension $G \times K$) is obtained by combining $\hat{\boldsymbol{\beta}}_g$ and $\hat{\boldsymbol{u}}_g$ (fixed effect + random effect), with respect to each corresponding cell type and subject condition. Elements in $\boldsymbol{R}_n$ could be computed as: $r_{gnk} = \hat{m}_{k,g} + z_n\hat{\beta}_{k,g} + \hat{u}_{n,k,g}$.

2.3 Stage III: Personalized deconvolution

With the subject-specific reference panel $\boldsymbol{R}_n$ and the original bulk mixture transcriptome data, as shown in the bottom-right of **Figure 1**, we use *non-negative least square* [32, 31] to deconvolute the cell type abundance $\boldsymbol{\Theta}_{I,n}$. $\boldsymbol{\Theta}_{I,n}$ is of dimension $K \times t_n$ for each subject respectively and the I in the subscript stands for the ***imply***-estimated cell type abundance. To be specific, we optimize the following objective function in *non-negative least square*:

$$\text{argmin}\|\boldsymbol{R}_n \otimes \boldsymbol{I}_{tn} vec(\boldsymbol{\Theta}'_{I,n}) - vec(\boldsymbol{Y}'_{\cdot n\cdot})\|_2$$

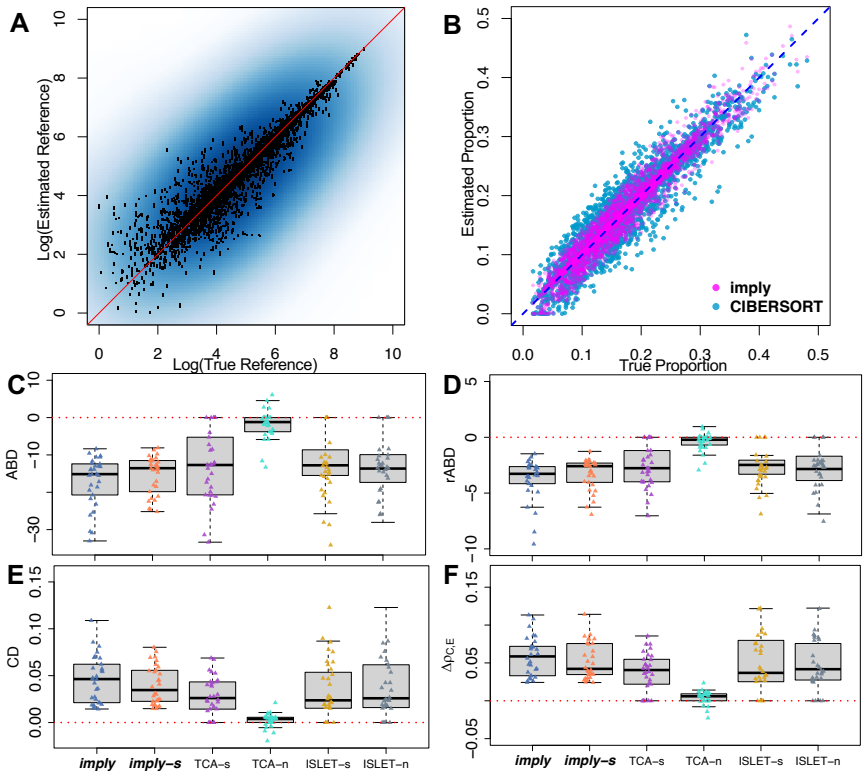

Fig. 2. ***imply*** improves cell type deconvolution accuracy. **(A)** Scatterplot showing ***imply*** estimated gene expression reference panel versus the true reference panel. **(B)** Superimposed scatterplot of the ***imply***-estimated cell type proportion over the CIBERSORT-estimates (the current state-of-art method). ***imply*** shows better concordance with the ground truth. **(C)**-**(F)** Boxplots displaying evaluation metrics and each point representing one simulation iteration: *ABD*, *rABD*, *CD*, and $\Delta\rho_{\mathrm{C,E}}$. Five additional modeling frameworks are benchmarked. The red dashed line (value of 0) represents no improvement in proportion estimation. For **(C)** and **(D)**, lower values indicate better deconvolution accuracy. For **(E)** and **(F)**, higher the better.

under the constraint $\boldsymbol{\Theta}_{I,n} \geq 0$. $\boldsymbol{Y}_{\cdot n \cdot}$ is the mixture data for subject n with dimension $G \times t_n$. This is a joint optimization across all the samples per subject simultaneously instead of sample-wise optimization, using the subject-specific $\boldsymbol{R}_n$ and quadratic programming. Overall, instead of using the population-level signature matrix $\boldsymbol{E}$, the adoption of personalized $\boldsymbol{R}_n$'s, for more genes, would benefit cell type abundance inferences.

2.4 Simulations

Pure cell-type-specific expression profiles The simulation scheme is adapted from a benchmark study [39] based on true cell line RNA-seq datasets [35]. The variance-covariance matrices and mean vectors of cell-type-specific gene expression means ($\hat{\boldsymbol{\Sigma}}_m$, $\bar{\boldsymbol{\mu}}_m$) and biological dispersion ($\hat{\boldsymbol{\Sigma}}_\phi$, $\bar{\boldsymbol{\mu}}_\phi$) are estimated by *PROPER* [56]. We use *Multivariate Normal Distribution* (MVN) to simulate expression mean ($\boldsymbol{M} \sim MVN(\bar{\boldsymbol{\mu}}_m, \hat{\boldsymbol{\Sigma}}_m)$) and dispersion ($\boldsymbol{\Phi} \sim MVN(\bar{\boldsymbol{\mu}}_\phi, \hat{\boldsymbol{\Sigma}}_\phi)$). The effect size of differential expression is defined as Log-Fold-Change (LFC) denoted as Δ. The true cell-type-specific gene expression matrix P is derived from *Gamma Distribution*: $\boldsymbol{P}_{case/ctrl} \sim \Gamma(\frac{1}{\exp(\Phi)}, \exp(\boldsymbol{M}_{case/ctrl}) \times \exp(\boldsymbol{\Phi}))$, where $\boldsymbol{M}_{ctrl} = \boldsymbol{M}, \boldsymbol{M}_{case} = \boldsymbol{M} + \Delta$.

Subject-to-subject variations (SSV) are added to $\boldsymbol{P}_{case/ctrl}$ to obtain subject-specific underlying gene expression matrices $\boldsymbol{P}_n$. SSV ranges from 0-5%, up to 20%-50%. $\boldsymbol{P}_n$ is shared across three simulated samples per subject.

Cell type proportions and observed read counts To generate the cell type proportions, we use *Dirichlet Distribution* to estimate $\boldsymbol{\alpha}$ parameters from multiple well-labeled single cell RNA-seq studies and then simulate cell type proportions: $\boldsymbol{\theta}_{\boldsymbol{T},\boldsymbol{ni}\cdot} \sim Dirichlet(\boldsymbol{\alpha}_{ctrl/case})$. $\boldsymbol{\theta}_{\boldsymbol{T},\boldsymbol{ni}\cdot}$ are reorganized into cell composition matrix, $\boldsymbol{\Theta}_T$. The sample-specific underlying gene expression panel is computed as $\boldsymbol{\lambda}_{ni} = \boldsymbol{P}_n \times \boldsymbol{\theta}'_{T,ni\cdot}$, and

follows a *Gamma Distribution* as well [41]. $\boldsymbol{\lambda}_{ni}$ is further assessed by the *Poisson Distribution* to generate observed RNA-seq counts data across the entire genome: $\boldsymbol{y}_{ni} \sim Pois(\boldsymbol{\lambda}_{ni})$.

Overall, the *Gamma Distribution* models biological variations, the *Dirichlet Distribution* regulates cell proportion, and *Poisson Distribution* mimics technical noise in the sequencing experiments.

2.5 Evaluation Metrics

We denote ***imply***'s deconvolution values as $\hat{\boldsymbol{\Theta}}_I$, the existing method's deconvolution results as $\hat{\boldsymbol{\Theta}}_E$, and the ground truth as $\boldsymbol{\Theta}$. The central goal is to assess how much improvement in cell proportion estimation ***imply*** could achieve. The following evaluation metrics are adopted for benchmarking:

Absolute bias differences (ABD) and relative absolute bias differences (rABD):

$$ABD := \sum |\hat{\boldsymbol{\Theta}}_I - \boldsymbol{\Theta}| - \sum |\hat{\boldsymbol{\Theta}}_E - \boldsymbol{\Theta}|; \qquad rABD := [\boldsymbol{Avg}(\frac{|\hat{\boldsymbol{\Theta}}_I - \boldsymbol{\Theta}|}{\boldsymbol{\Theta}}) - \boldsymbol{Avg}(\frac{|\hat{\boldsymbol{\Theta}}_E - \boldsymbol{\Theta}|}{\boldsymbol{\Theta}})] \times 100\%$$

For both ABD and $rABD$, if they are smaller than zero, it means ***imply*** successfully reduces the estimation bias. A smaller value further indicates better performance.

Correlation differences (CD):

$$CD := corr(\hat{\boldsymbol{\Theta}}_I, \boldsymbol{\Theta}) - corr(\hat{\boldsymbol{\Theta}}_E, \boldsymbol{\Theta})$$

$CD>0$ indicates ***imply*** increases the correlation between the estimation and the ground truth. A larger value indicates favorable performance.

Lin's concordance correlation coefficient (Lin's CCC) and its variations: Lin's CCC has been extensively used to evaluate the concordance between estimated measurements and gold standards [30]:

$$\rho_{\mathrm{C}}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}) = 1 - \frac{E[(\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}})^2]}{E_I[(\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}})^2]},$$

where E_I indicates the expectation under the assumption that $\boldsymbol{\Theta}$ and $\hat{\boldsymbol{\Theta}}$ are independent. Lin's CCC is bounded between 1 (perfect agreement) and -1 (disagreement). The concordance improves as $\rho_{\mathrm{C}}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}})$ approaches 1. We adopt a Euclidean distance-based variation of Lin's CCC, by substituting the expected squared difference to Euclidean distance, defined as $\rho_{\mathrm{C,E}}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}) = 1 - \frac{E[\sum_{k=1}^{K}(\boldsymbol{\Theta}^{(k)} - \hat{\boldsymbol{\Theta}}^{(k)})^2]}{E_I[\sum_{k=1}^{K}(\boldsymbol{\Theta}^{(k)} - \hat{\boldsymbol{\Theta}}^{(k)})^2]}$. Aitchison [2] distance-based Concordance Correlation Coefficient (CCC) is shown in supplementary section 1.3 and 3.6. These metrics are more statistically rigorous for compositional outcomes that are subject to the positiveness and unit-sum constraints [13]. If ***imply*** yields increased concordance and improved precision, we expect positive values in the differences of CCC, respectively defined as:

$$\Delta\rho_{\mathrm{C}} = \rho_{\mathrm{C}}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}_I) - \rho_{\mathrm{C}}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}_E); \qquad \Delta\rho_{\mathrm{C,E}} = \rho_{\mathrm{C,E}}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}_I) - \rho_{\mathrm{C,E}}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}_E)$$

3 RESULTS

We first evaluate ***imply***'s deconvolution accuracy using synthetic data. ***imply*** is the only method that re-estimates cell type proportions using personalized reference panels from longitudinal bulk data; therefore, a direct comparison with existing methods is not directly available. Nevertheless, we designed the benchmark to be inclusive of comparable methods. TCA [47], designed for csDE genes detection, integrates a cell proportion re-estimation feature. TCA takes a maximum-likelihood (ML) approach to derive model parameters, and proportions are subsequently updated. Since TCA requires preliminary cell proportions for re-estimation, we employ *non-negative least squares* and ν-SVR to acquire the initial inputs and label them as TCA-n and TCA-s. ISLET [16] is the first method to retrieve individual-specific reference estimation in repeated samples based on the Expectation-Maximization (EM) algorithm. ISLET can be an alternative approach to our mixed-effect model to solve personalized reference panels. Here, we consider ISLET-s and ISLET-n to denote ISLET variants, with the final personalized deconvolution is conducted by SVR or *non-negative least squares*. We also introduce a variant of ***imply***, denoted as ***imply-s***, where where *Stage III* is achieved by SVR. We comprehensively benchmark our proposed methods, ***imply*** and its variant ***imply-s***, against other algorithms: TCA-n, TCA-s, ISLET-n, and ISLET-s.

3.1 *imply* increases precision in cell-type deconvolution

In the baseline scenario, we have 100 subjects per group, SSV up to 5%, and an effect size of 0.5. **Figure 2A** shows the ***imply***-estimated reference panels versus the ground truth. We observe good accuracy in personalized reference panel recovery, especially among high-expression genes. This demonstrates the fidelity of *Stage II* and lays a foundation for *Stage III*. Next, we evaluate if ***imply***'s cell type deconvolution from *Stage III* could reduce bias. **Figure 2B** shows the scatterplot of the estimated cell type proportions versus the true proportions. Our result is overlaid on top of the result from CIBERSORT. ***imply*** yields higher precision in deconvolution as its estimates aggregate closer to the diagonal line. In **Figure 2C-F**, the bias reductions are quantitatively assessed and compared by *ABD*, *rABD*, *CD*, and $\Delta\rho_{C,E}$. Each point in a boxplot represents an iteration. The zero line represents the existing deconvolution method, such as CIBERSORT, which did not consider personalized reference panels. For *ABD* and *rABD*, lower values indicate greater increases in deconvolution accuracy; while for *CD* and $\Delta\rho_{C,E}$, higher values indicate improved concordance with the ground truth. Notably, ***imply*** consistently demonstrates the most substantial reduction in deconvolution bias and highest improvement in concordance with the truth. In contrast, TCA performs poorly, especially when the initial proportion inputs are estimated through *non-negative least squares* (TCA-n). Even when the initial proportion input is derived from CIBERSORT, the bias reduction achieved by TCA (TCA-s) is not as significant as that achieved by ***imply***. Furthermore, we notice that personalized reference panels estimated by ISLET also yield benefits for personalized deconvolution, illustrated by ISLET-s and ISLET-n. However, the improvements are not as pronounced as those achieved by ***imply***. We also explore the methods' performance under various simulation scenarios. **Table 1** shows averaged *ABD*s across iterations, with each standard error. Bold fonts highlight the algorithm with the most bias reduction for each scenario. ***imply*** and ***imply-s*** consistently demonstrate exceptional performance in reducing deconvolution bias across all scenarios.

Table 1. Benchmarking ***imply*** across various simulation scenarios. The table shows *Absolute Bias Difference (ABD)* at various subject-specific variations (SSV), effect sizes (LFC), and sample sizes (N). *ABD* values are shown, along with their standard error in parentheses. A lower value indicates better deconvolution estimation improvement. The bold font indicates the best method in each scenario.

SSV	LFC	N	***imply***	***imply-s***	TCA-s	TCA-n	ISLET-s	ISLET-n
0%∼5%	0.5	25	**-3.95 (1.92)**	-3.51 (1.56)	-2.21 (1.96)	-0.33 (0.85)	1.21 (20.15)	1.03 (20.25)
		50	**-8.97 (3.86)**	-7.97 (3.44)	-6.6 (4.92)	-0.8 (2.2)	5.76 (52.49)	7.75 (57.2)
		100	**-17.08 (6.66)**	-15.38 (5.11)	-13.1 (9.43)	-1.88 (3.97)	25.71 (148.83)	15.11 (116.56)
	1	25	**-3.39 (2.98)**	-3.17 (2.43)	-0.89 (2.57)	-0.29 (0.64)	14.95 (40.81)	14.51 (36.8)
		50	**-8.59 (4.88)**	-7.63 (4.78)	-1.64 (4.03)	-1.11 (1.92)	45.7 (84.85)	35.33 (70.75)
		100	-16.49 (10.45)	**-16.54 (11.56)**	-4.17 (10.17)	-2.29 (3.78)	22.55 (121.26)	19.26 (115.44)
	1.25	25	**-4.9 (3.11)**	-4.28 (3.48)	-1.38 (2.75)	-0.35 (0.8)	4.55 (22.45)	2.22 (18.2)
		50	**-9.39 (5.56)**	-9.09 (8.35)	-4.47 (8.67)	-0.96 (1.79)	31.72 (80.43)	24.19 (73.19)
		100	-21.73 (13.28)	**-21.91 (20.18)**	-9.56 (16.92)	-2.13 (3.36)	52.94 (144.77)	30.96 (102.83)
5%∼10%	0.5	25	**-3.53 (1.36)**	-3.28 (1.18)	-1.88 (2.47)	-0.33 (1.52)	8.84 (37.92)	6.4 (31.77)
		50	**-7.6 (3.36)**	-6.79 (2.69)	-4.53 (5.13)	-0.51 (3.99)	2.88 (47.91)	-0.95 (29.74)
		100	**-15.64 (5.85)**	-15.42 (8.12)	-8.06 (11.59)	-1.9 (7.37)	7.15 (117.46)	5.51 (107.12)
	1	25	**-4.03 (2.45)**	-3.75 (2.39)	-0.22 (1.42)	0.06 (1.02)	15.35 (41.93)	16.46 (45.14)
		50	-7.93 (4.29)	**-8.01 (7.12)**	-0.45 (5.5)	-0.19 (2.94)	29.12 (79.78)	21.14 (68.21)
		100	**-15.86 (10.45)**	-13.97 (9.44)	-1.4 (11.86)	-0.35 (3.87)	57.17 (176.1)	58.45 (172.78)
	1.25	25	-4.21 (2.68)	**-6.55 (9.34)**	-0.65 (2.52)	-0.08 (0.61)	0.99 (23.14)	1.49 (13.97)
		50	**-9.85 (6.61)**	-8.83 (7.24)	-1.81 (6.96)	0.05 (2.35)	36.08 (83.23)	24.17 (63.91)
		100	-20.3 (12.05)	**-22.76 (24.38)**	-6.36 (18.9)	-1.44 (4.28)	28.14 (137.05)	21.45 (125.49)
10%∼20%	0.5	25	**-3.27 (1.64)**	-2.83 (1.34)	-0.1 (4.84)	0.65 (3.79)	13.74 (41.37)	13.22 (40.18)
		50	**-6.48 (2.86)**	-6.14 (2.27)	1.63 (15.12)	2.07 (10.22)	2.26 (38.29)	2.66 (40.29)
		100	**-13.98 (5.5)**	-12.73 (5.14)	1.32 (31.49)	9.86 (29.37)	7.05 (94.67)	7.02 (103.21)
	1	25	-2.6 (3.46)	**-3.49 (4.91)**	0.75 (3.1)	2.01 (4.04)	28.61 (44.79)	27.28 (45.55)
		50	**-7.75 (4)**	-7.43 (3.35)	1.86 (7.61)	5.18 (11.03)	19.09 (64.13)	17.76 (64.18)
		100	-14.52 (8.4)	**-14.8 (8.49)**	6.67 (19.22)	11.3 (26.41)	46.52 (124.95)	51.84 (140)
	1.25	25	**-4.77 (1.99)**	-4.22 (1.59)	0.41 (3.07)	1.98 (3.19)	12.72 (35.35)	6.78 (23.43)
		50	**-9.49 (4.57)**	-8.68 (4.45)	0.17 (6.14)	4.08 (7.69)	27.71 (73.1)	30.77 (84.76)
		100	**-19.13 (10.37)**	-16.49 (12.37)	0.89 (18.5)	8.69 (18.58)	-8.27 (16.34)	-9.42 (13.32)

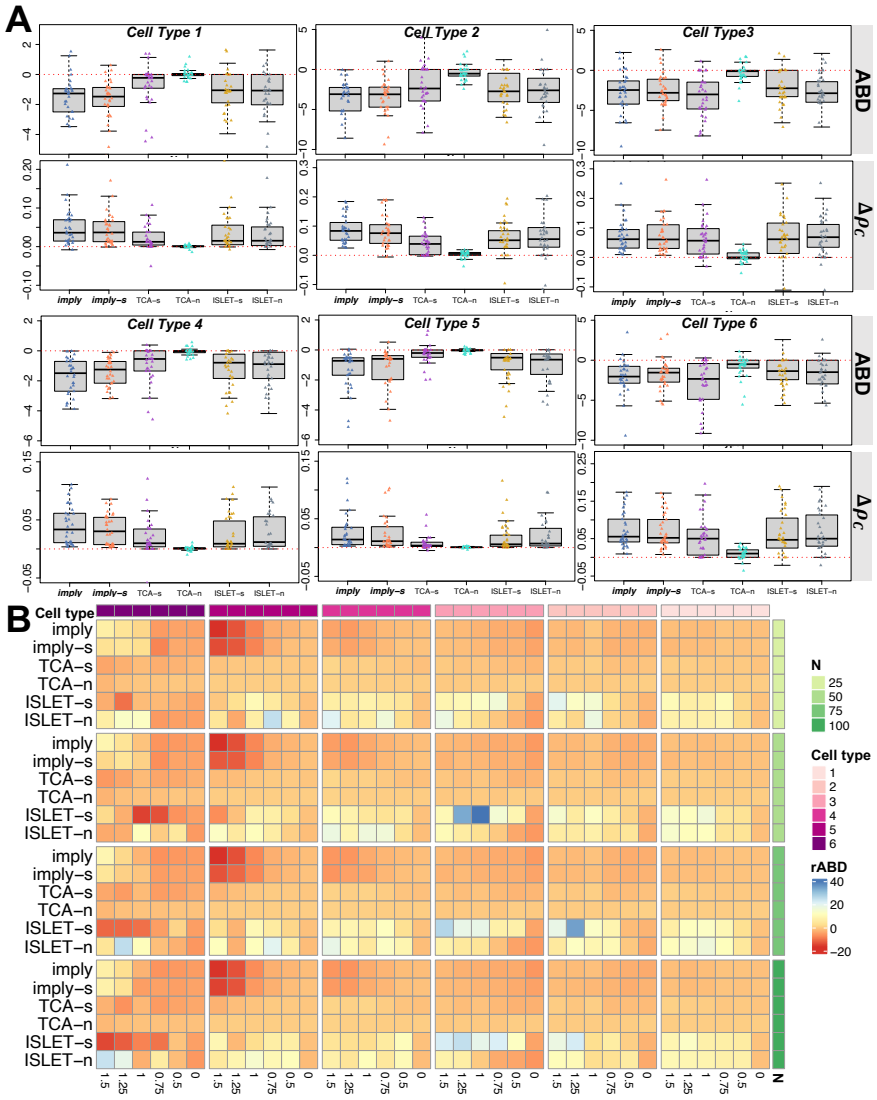

Fig. 3. Cell-type resolution improvements in proportions estimated by ***imply***. **(A)** Boxplots showing ABD and $\Delta\rho_{\mathrm{C}}$, for cell types 1 to 6. **(B)** Heatmap showing the deconvolution improvement using the $rABD$ metric, aggregated by cell types (top row) and sample sizes (right column), for various effect sizes (bottom row).

3.2 Benchmarking at cell-type resolution

We next investigate the deconvolution accuracy at cellular level. **Figure 3A** shows ABD and $\Delta\rho_{\mathrm{C}}$ of 30 replicates for each cell type when SSV, sample size, and effect size are set to 0-5%, 75, and 0.5. We can see a discernible reduction in bias when personalized reference panels are adopted. ***imply*** and ***imply-s*** consistently stand out, yielding a significant enhancement in concordance compared to others. The heatmap in **Figure 3B** shows the average $rABD$ at various combinations of sample sizes and effect sizes, separated by cell types. At large effect sizes, improvements in accuracies facilitated by ***imply*** are notably more profound. However, $rABD$ is insensitive to sample sizes. There is a connection between bias reduction and cell type abundances as shown in supplementary section 3: deconvolution accuracies for more abundant cells are highly sensitive to LFC changes. In contrast, for minor cell types, the small contribution amplifies deconvolution difficulties, as sequencing noise can overshadow biological variations.

3.3 Influential factors in deconvolution accuracy

We further zoom in to study how sample size, effect size, and SSV would affect personalized deconvolution. In **Figure 4A**, ABD and $\Delta\rho_{\mathrm{C,E}}$ for ***imply***, together with ISLET-n and TCA-s, are presented across LFC

ranging from 0 (null) to 1.5. ***imply*** consistently exhibits the lowest ABD in all scenarios and the highest $\Delta\rho_{\mathrm{C,E}}$ in most settings. These results indicate the advantage of adopting personalized reference panels. In addition, ***imply*** is the most stable (i.e., smallest variation) as the effect size increases. **Figure 4B** shows the same metrics across various sample sizes. As expected, ABD decreases as the sample size increases. ***imply*** consistently maintains the highest $\Delta\rho_{\mathrm{C,E}}$ across various sample sizes. In **Figure 4C**, we further investigate the $\Delta\rho_{\mathrm{C,E}}$ alteration percentages, which are defined as $\Delta\rho_{\mathrm{C,E}}\% = \frac{\Delta\rho_{\mathrm{C,E}}}{\rho_{\mathrm{C,E}}(\Theta_E,\Theta)} \times 100\%$, at different levels of SSV. We observe a robust pattern across different effect sizes, samples sizes, and SSVs, and conclude that ***imply*** and ***imply-s*** consistently provide the most outstanding concordance improvement.

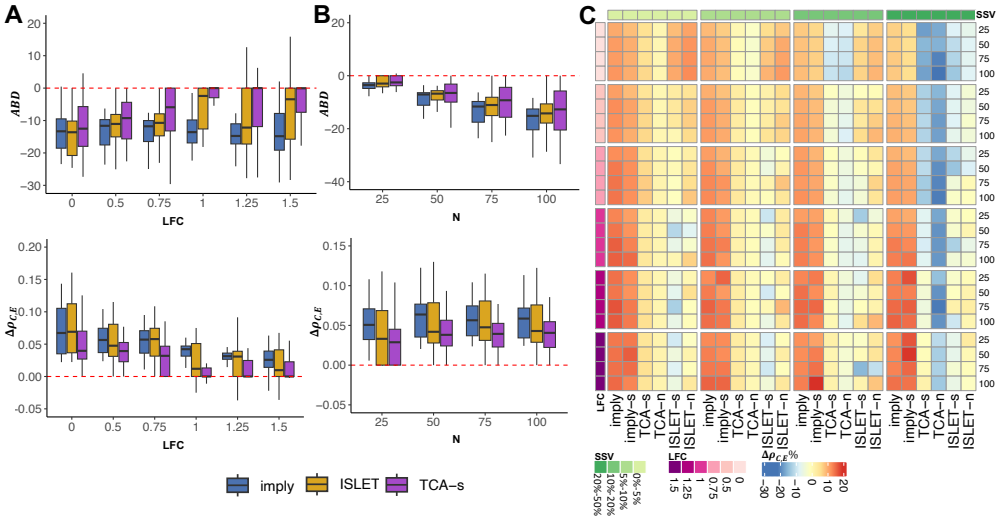

Fig. 4. Effect size, sample size, and SSV affect deconvolution accuracy. **(A)** Boxplots of ABD (upper panel) and $\Delta\rho_{\mathrm{C,E}}$ (lower panel) across three methods: ***imply***, ISLET-n, TCA-s, under different effect sizes (LFC). **(B)** Similar to **(A)** but across various sample sizes per group. **(C)** Heatmap showing the relative $\Delta\rho_{\mathrm{C,E}}$ across various combinations of sample sizes, effect sizes, and SSV. The color bars on the left and the top indicate the LFC and SSV, respectively. The number on the right indicates the sample size per group.

3.4 Application of *imply* to longitudinal transcriptomic datasets

We applied ***imply*** to two consortia longitudinal transcriptomic datasets: Parkinson's Disease Biomarker Program (PDBP) and The Environmental Determinants of Diabetes in the Young (TEDDY). The PDBP consortium has the longitudinal RNA-seq dataset extracted from the whole blood. De-identified participants with at least three observations over time were retained. A total of 399 PD patients and 173 controls, with 2599 longitudinal samples over 2 years, were included. Clinical data includes information about patients' medical history, symptoms, disease status, total Montreal Cognitive Assessment (MoCA) scores, and MDS UPDRS part III motor scores. The TEDDY cohort is a multi-center pediatric study of Type 1 Diabetes (T1D). TEDDY cohort screened and enrolled participants with susceptibility of T1D based on the Human Leukocyte Antigen (HLA) genotypes from six clinical centers in four countries. A total of 8,676 high-risk infants were enrolled from birth and followed every 3 months for blood sample collection and islet autoantibody (IAbs) measurement up to 4 years of age. Details of sample collection, RNA sequencing procedures, and quality control in TEDDY are described in [57].

Figure 5 shows the deconvolution analysis results for PDBP and TEDDY. For PDBP, the mean proportions across all visit times of six cell types are shown for cases and controls in **Figure 5A**. Here, B cell contributes the most among all six cell types, while NK cell contributes the least. We notice a higher CD8 cell proportions in the PD group than in the control group, while CD4 cell proportions in the PD groups are lower. **Figure 5B** displays the heatmap of Pearson correlations among the six cell types. B cells, monocytes, and CD4 all show negative pairwise correlations. **Figure 5C** shows boxplots of CD8 cell proportions

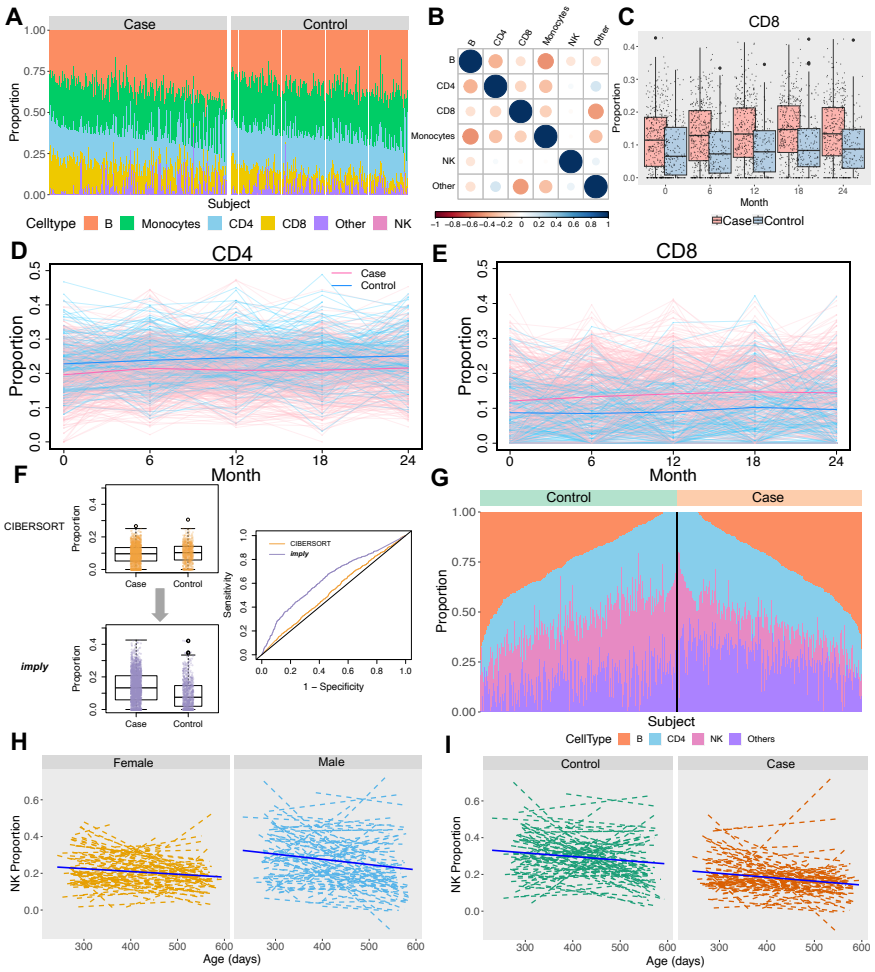

Fig. 5. Phenotype-associated cell type disparities from PDBP (Parkinson's Disease Biomarker Program) and TEDDY (The Environmental Determinants of Diabetes in the Young) consortia. **(A)**: Cell type proportions for all subjects, separated by PD status in PDBP dataset. The bar represents the mean cell type proportions across all visit times, for each subject. **(B)**: Pearson correlations between six cell type proportions among all individuals. **(C)**: Distribution comparisons of CD8 proportions between PD and controls, at each visit time. **(D)**: Deconvoluted CD4 proportions along with study participants' visit time. PD (pink) and controls (blue) are illustrated by both individual background lines (thin) and foreground lines (thick). **(E)**: Same as in **(D)** but for CD8 proportions. **(F)**: Grand comparison of CD8 proportions between PD and controls, using CIBERSORT and ***imply***. Results from ***imply*** show a larger effect size, more significant test statistics, and increased discriminative capacity. **(G)**: Cell type proportions for all subjects, separated by pancreatic islet autoantibodies (IA) status, in TEDDY dataset. The bar represents the mean cell type proportions across all visit times, for each subject. **(H)**: NK proportions along infant's age (in days) at sample collection, for female and male subjects. Average fitted lines (solid) overlay individual-specific lines (dashed). **(I)**: Same as in **(H)** but separated by IA case and control status.

comparing case and control, at each time point. The median value of CD8 proportion in case is higher than that in control group at each time point. The CD4 and CD8 cell type proportions, broken down by the participant's visit time of each subject, are shown in **Figure 5D** and **5E**, respectively. For CD4, the mean proportions in case are lower than those in control for each visit time. For CD8, the mean proportions among cases are higher than those among controls, for each visit time. These findings are well-aligned with previous studies where the PD patients showed elevated CD8 proportions and reduced CD4 proportions than controls [54, 7, 19]. We also benchmarked ***imply*** with CIBERSORT as shown in **Figure 5F**. Using CIBERSORT, the p-value of the Wilcoxon Rank Sum test is 0.0111 and the median difference is -0.007 for CD8 proportions between cases and controls. It incorrectly suggests that the CD8 proportion of cases is lower than

controls. In contrast, ***imply*** yields a p-value less than 10^{-16} and the median difference is 0.58, which shows the correct effect size direction. It also increases differential power between cases and controls, as shown in the ROC plot. Associations between the various cell type proportions and clinical outcomes, including total UPSIT score, total scores of MoCA, Cerebrospinal fluid (CSF), and MDS UPDRS part III motor scores, are detailed in supplementary section 4.1. For the T1D study of TEDDY, the disease status of interest is the onset of pancreatic islet autoantibodies (IA). The longitudinal analysis of re-quantified cellular composition identifies NK cell abundance as higher in males than females ($p < 0.0001$), as illustrated in **Figure 5H**. Previous research in TEDDY reported a higher risk of IA being associated with viral infection during the first 6 months of life [53]. The sex difference in NK cell fraction in **Figure 5H** could be a consequence of early-life vaccination or viral infection [10], since infants are exposed to exogenous antigens and have a high susceptibility to infections. In this analysis, we use longitudinal samples of IA cases and controls collected at the age of 9-21 months, and compare deconvoluted cell fractions between groups by ***imply***. **Figure 5I** shows that the NK cell proportions are significantly lower ($p < 0.0001$) in the participants who developed IA at a young age compared to controls, while this trend is not observed in the initial cell abundance estimated by CIBERSORT ($p = 0.77$, supplementary section 4.2). The relative higher NK cell abundance in males (vs. females) and controls (vs. cases) among TEDDY participants is consistent with the previous finding that males have a lower risk of autoimmunity than females [38].

Furthermore, we perform a downstream csDE genes analysis on IA status based on ***imply***-deconvoluted cell type fraction, using ISLET [16] with FDR< 0.1. The cell type proportions improved by ***imply*** enabled the detection of DE genes in CD4 T cells and identified more NK-cell-specific DE genes ($n > 300$) compared to a previous csDE genes testing result ($n = 30$) based on the proportions deconvoluted by AutoGeneS [3]. The IA-csDE genes based on the improved cell fractions include the markers for multiple T cell receptors (e.g., TRBV, TRDV, TRGV, TRJV) and the genes regulating immune responses such as *CAMP* and *CRK*. The *CAMP* gene expression was found to be associated with serum levels of vitamin D in the studies of innate immunity [37, 23, 45], while the TEDDY cohort also reported a strong linkage between vitamin D and the risk of IA [33]. Protein *CRK* is involved in NK cells inhibitory receptor signaling and modulates the signaling of activating receptors, which may function as a two-way molecular switch to control NK cell-mediated cytotoxicity [42, 36].

4 Discussion

The computational deconvolution of admixed bulk tissue samples is drawing substantial interest as large consortia are becoming increasingly available. We are among the first to consider personalized reference panels in deconvolution. ***imply*** optimizes the usages of shared information in longitudinal samples from each subject and jointly quantifies the cell abundances across multiple samples per subject. We show the advantage of using personalized reference panels by *in silico* simulation studies and the analytical results of two large-scale longitudinal consortia. ***imply*** can produce more accurate and realistic results. Alternative machine learning approaches, such as EM and non-negative matrix factorization algorithms, could also extract personalized reference panels and have been implemented in ISLET [16] and CIBERSORTx [44]. Nevertheless, they lack the conciseness and computational efficiency exhibited by the proposed linear mixed-effects modeling framework.

A limitation of ***imply*** is the requirement of an initial signature matrix as the input in *Stage I*, which could affect the initial cell type abundance estimation as the input for downstream. An alternative approach is to initialize cell fractions by external multi-subject reference cell count data, such as single-cell profiling and labeling, flow cytometry, or imaging. For some genes, the random effect variance estimation may shrink towards zero, likely due to the adoption of penalized MLE. For such scenarios, the cell-type-specific heterogeneity between individuals would not be fully recovered. Furthermore, the intra-individual heterogeneity was not considered in reference panel recovery. This is because our present work was motivated by the bulk transcriptome of longitudinal blood samples, many of which were collected from healthy controls. In those scenarios, the underlying pure gene expression panel for each subject is relatively stable over time. Our previous work [16] suggests that the intra-individual cell-type-specific heterogeneity, when assessing using longitudinal PBMC scRNA-seq data, is trivial when compared with inter-individual variation. Hence, our future work will include the curation of longitudinal scRNA-seq data from distinct tissue types or disease populations and the incorporation of potential variations between time points at cell type resolution.

References

1. Aguirre-Gamboa, R., Joosten, I., Urbano, P.C., van der Molen, R.G., van Rijssen, E., van Cranenbroek, B., Oosting, M., Smeekens, S., Jaeger, M., Zorro, M., et al.: Differential effects of environmental and genetic factors on t and b cell immune traits. Cell reports **17**(9), 2474–2487 (2016)
2. Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A., Pawlowsky-Glahn, V.: Logratio analysis and compositional distance. Mathematical geology **32**, 271–275 (2000)
3. Aliee, H., Theis, F.J.: Autogenes: automatic gene selection using multi-objective optimization for rna-seq deconvolution. Cell Systems **12**(7), 706–715 (2021)
4. Aran, D., Hu, Z., Butte, A.J.: xcell: digitally portraying the tissue cellular heterogeneity landscape. Genome biology **18**, 1–14 (2017)
5. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J.E., Mestdagh, P., De Preter, K.: Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nature communications **11**(1), 5650 (2020)
6. Avila Cobos, F., Vandesompele, J., Mestdagh, P., De Preter, K.: Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics **34**(11), 1969–1979 (2018)
7. Baba, Y., Kuroiwa, A., Uitti, R.J., Wszolek, Z.K., Yamada, T.: Alterations of t-lymphocyte populations in parkinson disease. Parkinsonism & related disorders **11**(8), 493–498 (2005)
8. Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823 (2014)
9. Çalışkan, M., Baker, S.W., Gilad, Y., Ober, C.: Host genetic variation influences gene expression response to rhinovirus infection. PLoS genetics **11**(4), e1005111 (2015)
10. Cheng, M.I., Li, J.H., Riggan, L., Chen, B., Tafti, R.Y., Chin, S., Ma, F., Pellegrini, M., Hrncir, H., Arnold, A.P., et al.: The x-linked epigenetic regulator utx controls nk cell-intrinsic sex differences. Nature Immunology pp. 1–12 (2023)
11. Chiu, Y.J., Hsieh, Y.H., Huang, Y.H.: Improved cell composition deconvolution method of bulk gene expression profiles to quantify subsets of immune cells. BMC medical genomics **12**, 1–17 (2019)
12. Clarke, J., Seo, P., Clarke, B.: Statistical expression deconvolution from mixed tissue samples. Bioinformatics **26**(8), 1043–1049 (2010)
13. Cui, Y., Peng, L., Hu, Y., Lai, H.J.: Assessing the reproducibility of microbiome measurements based on concordance correlation coefficients. Journal of the Royal Statistical Society Series C: Applied Statistics **70**(4), 1027–1048 (2021)
14. Di Biase, M.A., Geaghan, M.P., Reay, W.R., Seidlitz, J., Weickert, C.S., Pébay, A., Green, M.J., Quidé, Y., Atkins, J.R., Coleman, M.J., et al.: Cell type-specific manifestations of cortical thickness heterogeneity in schizophrenia. Molecular psychiatry **27**(4), 2052–2060 (2022)
15. Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C.M., Zou, F., Jiang, Y.: Scdc: bulk gene expression deconvolution by multiple single-cell rna sequencing references. Briefings in bioinformatics **22**(1), 416–427 (2021)
16. Feng, H., Meng, G., Lin, T., Parikh, H., Pan, Y., Li, Z., Krischer, J., Li, Q.: Islet: individual-specific reference panel recovery improves cell-type-specific inference. Genome Biology **24**(1), 174 (2023)
17. Findley, A.S., Monziani, A., Richards, A.L., Rhodes, K., Ward, M.C., Kalita, C.A., Alazizi, A., Pazokitoroudi, A., Sankararaman, S., Wen, X., et al.: Functional dynamic genetic effects on gene regulation are specific to particular cell types and environmental conditions. Elife **10**, e67077 (2021)
18. Finotello, F., Trajanoski, Z.: Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunology, Immunotherapy **67**(7), 1031–1040 (2018)
19. Galiano-Landeira, J., Torra, A., Vila, M., Bove, J.: Cd8 t cell nigral infiltration precedes synucleinopathy in early stages of parkinson's disease. Brain **143**(12), 3717–3733 (2020)
20. Gaujoux, R., Seoighe, C.: Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. Infection, Genetics and Evolution **12**(5), 913–921 (2012)
21. Ghaffari, S., Bouchonville, K.J., Saleh, E., Schmidt, R.E., Offer, S.M., Sinha, S.: Bedwars: a robust bayesian approach to bulk gene expression deconvolution with noisy reference signatures. Genome Biology **24**(1), 1–30 (2023)
22. Gibson, G.: The environmental contribution to gene expression profiles. Nature reviews genetics **9**(8), 575–581 (2008)
23. Gombart, A.F., Saito, T., Koeffler, H.P.: Exaptation of an ancient alu short interspersed element provides a highly conserved vitamin d-mediated innate immune response in humans and primates. BMC genomics **10**(1), 1–11 (2009)
24. Gong, T., Szustakowski, J.D.: Deconrnaseq: a statistical framework for deconvolution of heterogeneous tissue samples based on mrna-seq data. Bioinformatics **29**(8), 1083–1085 (2013)
25. Huang, P., Cai, M., Lu, X., McKennan, C., Wang, J.: Accurate estimation of rare cell type fractions from tissue omics data via hierarchical deconvolution. bioRxiv pp. 2023–03 (2023)

26. Idaghdour, Y., Czika, W., Shianna, K.V., Lee, S.H., Visscher, P.M., Martin, H.C., Miclaus, K., Jadallah, S.J., Goldstein, D.B., Wolfinger, R.D., et al.: Geographical genomics of human leukocyte gene expression variation in southern morocco. Nature genetics **42**(1), 62–67 (2010)
27. Kedlian, V.R., Donertas, H.M., Thornton, J.M.: The widespread increase in inter-individual variability of gene expression in the human brain with age. Aging (Albany NY) **11**(8), 2253 (2019)
28. Kuhn, A., Kumar, A., Beilina, A., Dillman, A., Cookson, M.R., Singleton, A.B.: Cell population-specific expression analysis of human cerebellum. BMC genomics **13**, 1–15 (2012)
29. Kuhn, A., Thu, D., Waldvogel, H.J., Faull, R.L., Luthi-Carter, R.: Population-specific expression analysis (psea) reveals molecular changes in diseased brain. Nature methods **8**(11), 945–947 (2011)
30. Lawrence, I., Lin, K.: A concordance correlation coefficient to evaluate reproducibility. Biometrics pp. 255–268 (1989)
31. Lawson, C.L., Hanson, R.J.: Solving least squares problems. prentice-hall inc., englewood cliffs, new jersey, p. 263 (1974)
32. Lawson, C.L., Hanson, R.J.: Solving least squares problems. SIAM (1995)
33. Li, Q., Liu, X., Yang, J., Erlund, I., Lernmark, Å., Hagopian, W., Rewers, M., She, J.X., Toppari, J., Ziegler, A.G., et al.: Plasma metabolome and circulating vitamins stratified onset age of an initial islet autoantibody and progression to type 1 diabetes: the teddy study. Diabetes **70**(1), 282–292 (2021)
34. Li, Z., Wu, Z., Jin, P., Wu, H.: Dissecting differential signals in high-throughput data from complex tissues. Bioinformatics **35**(20), 3898–3905 (2019)
35. Linsley, P.S., Speake, C., Whalen, E., Chaussabel, D.: 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 (2014)
36. Liu, D.: The adaptor protein crk in immune response. Immunology and cell biology **92**(1), 80–89 (2014)
37. Lowry, M.B., Guo, C., Zhang, Y., Fantacone, M.L., Logan, I.E., Campbell, Y., Zhang, W., Le, M., Indra, A.K., Ganguli-Indra, G., et al.: A mouse model for vitamin d-induced human cathelicidin antimicrobial peptide gene expression. The Journal of steroid biochemistry and molecular biology **198**, 105552 (2020)
38. Markle, J.G., Frank, D.N., Mortin-Toth, S., Robertson, C.E., Feazel, L.M., Rolle-Kampczyk, U., Von Bergen, M., McCoy, K.D., Macpherson, A.J., Danska, J.S.: Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. Science **339**(6123), 1084–1088 (2013)
39. Meng, G., Tang, W., Huang, E., Li, Z., Feng, H.: A comprehensive assessment of cell type-specific differential expression methods in bulk data. Briefings in bioinformatics **24**(1), bbac516 (2023)
40. Modlich, O., Prisack, H.B., Munnes, M., Audretsch, W., Bojar, H.: Immediate gene expression changes after the first course of neoadjuvant chemotherapy in patients with primary breast cancer disease. Clinical cancer research **10**(19), 6418–6431 (2004)
41. Moschopoulos, P.G.: The distribution of the sum of independent gamma random variables. Annals of the Institute of Statistical Mathematics **37**(1), 541–544 (1985)
42. Nabekura, T., Chen, Z., Schroeder, C., Park, T., Vivier, E., Lanier, L.L., Liu, D.: Crk adaptor proteins regulate nk cell expansion and differentiation during mouse cytomegalovirus infection. The Journal of Immunology **200**(10), 3420–3428 (2018)
43. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., Alizadeh, A.A.: Robust enumeration of cell subsets from tissue expression profiles. Nature methods **12**(5), 453–457 (2015)
44. Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al.: Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature biotechnology **37**(7), 773–782 (2019)
45. de Oliveira, A.L.G., Chaves, A.T., Cardoso, M.S., Pinheiro, G.R.G., Antunes, D.E., de Faria Grossi, M.A., Lyon, S., Bueno, L.L., da Costa Rocha, M.O., da Silva Menezes, C.A., et al.: Reduced vitamin d receptor (vdr) and cathelicidin antimicrobial peptide (camp) gene expression contribute to the maintenance of inflammatory immune response in leprosy patients. Microbes and Infection **24**(6-7), 104981 (2022)
46. Qiao, W., Quon, G., Csaszar, E., Yu, M., Morris, Q., Zandstra, P.W.: Pert: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. PLoS computational biology **8**(12), e1002838 (2012)
47. Rahmani, E., Schweiger, R., Rhead, B., Criswell, L.A., Barcellos, L.F., Eskin, E., Rosset, S., Sankararaman, S., Halperin, E.: Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. Nature communications **10**(1), 3417 (2019)
48. Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L.: New support vector algorithms. Neural computation **12**(5), 1207–1245 (2000)
49. Shen-Orr, S.S., Tibshirani, R., Khatri, P., Bodian, D.L., Staedtler, F., Perry, N.M., Hastie, T., Sarwal, M.M., Davis, M.M., Butte, A.J.: Cell type–specific gene expression differences in complex tissues. Nature methods **7**(4), 287–289 (2010)

50. Sutton, G.J., Poppe, D., Simmons, R.K., Walsh, K., Nawaz, U., Lister, R., Gagnon-Bartsch, J.A., Voineagu, I.: Comprehensive evaluation of deconvolution methods for human brain gene expression. Nature Communications **13**(1), 1358 (2022)
51. Troester, M.A., Hoadley, K.A., Sørlie, T., Herbert, B.S., Børresen-Dale, A.L., Lønning, P.E., Shay, J.W., Kaufmann, W.K., Perou, C.M.: Cell-type-specific responses to chemotherapeutics in breast cancer. Cancer research **64**(12), 4218–4226 (2004)
52. Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., Yuan, G.C.: Accurate estimation of cell-type composition from gene expression data. Nature communications **10**(1), 2975 (2019)
53. Vehik, K., Lynch, K.F., Wong, M.C., Tian, X., Ross, M.C., Gibbs, R.A., Ajami, N.J., Petrosino, J.F., Rewers, M., Toppari, J., et al.: Prospective virome analyses in young children at increased genetic risk for type 1 diabetes. Nature medicine **25**(12), 1865–1872 (2019)
54. Wang, P., Yao, L., Luo, M., Zhou, W., Jin, X., Xu, Z., Yan, S., Li, Y., Xu, C., Cheng, R., et al.: Single-cell transcriptome and tcr profiling reveal activated and expanded t cell populations in parkinson's disease. Cell Discovery **7**(1), 52 (2021)
55. Wang, X., Park, J., Susztak, K., Zhang, N.R., Li, M.: Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature communications **10**(1), 380 (2019)
56. Wu, H., Wang, C., Wu, Z.: Proper: comprehensive power evaluation for differential expression using rna-seq. Bioinformatics **31**(2), 233–241 (2015)
57. Xhonneux, L.P., Knight, O., Lernmark, Å., Bonifacio, E., Hagopian, W.A., Rewers, M.J., She, J.X., Toppari, J., Parikh, H., Smith, K.G., et al.: Transcriptional networks in at-risk individuals identify signatures of type 1 diabetes progression. Science translational medicine **13**(587), eabd5666 (2021)
58. Zhong, Y., Wan, Y.W., Pang, K., Chow, L.M., Liu, Z.: Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC bioinformatics **14**(1), 1–10 (2013)