01 Jun 2022

Robust and Accurate Estimation of Cellular Fraction from Tissue Omics Data Via Ensemble Deconvolution

Manqi Cai

Molin Yue

Tianmeng Chen

Jinling Liu
Missouri University of Science and Technology, jinling.liu@mst.edu

et. al. For a complete list of authors, see https://scholarsmine.mst.edu/engman_syseng_facwork/881

Recommended Citation

M. Cai and M. Yue and T. Chen and J. Liu and E. Forno and X. Lu and T. Billiar and J. Celedón and C. Mckennan and W. Chen and J. Wang, "Robust and Accurate Estimation of Cellular Fraction from Tissue Omics Data Via Ensemble Deconvolution," *Bioinformatics*, vol. 38, no. 11, pp. 3004 - 3010, Oxford University Press, Jun 2022.
The definitive version is available at https://doi.org/10.1093/bioinformatics/btac279

Bioinformatics, 38(11), 2022, 3004–3010
https://doi.org/10.1093/bioinformatics/btac279
Advance Access Publication Date: 19 April 2022
Original Paper

Gene expression

Robust and accurate estimation of cellular fraction from tissue omics data via ensemble deconvolution

Manqi Cai[1], Molin Yue[1], Tianmeng Chen[2,3], Jinling Liu[4,5], Erick Forno[6], Xinghua Lu [7], Timothy Billiar[2], Juan Celedón [6], Chris McKennan[8], Wei Chen [6,*] and Jiebiao Wang [1,*]

[1]Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA, [2]Department of Surgery, University of Pittsburgh, Pittsburgh, PA 15213, USA, [3]Department of Pathology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA, [4]Department of Engineering Management and Systems Engineering, Missouri University of Science and Technology, Rolla, MO 65409, USA, [5]Department of Biological Sciences, Missouri University of Science and Technology, Rolla, MO 65409, USA, [6]Department of Pediatrics, University of Pittsburgh Medical Center Children's Hospital of Pittsburgh, Pittsburgh, PA 15224, USA, [7]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15206, USA and [8]Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15213, USA

*To whom correspondence should be addressed.
Associate Editor: Inanc Birol

Received on January 10, 2022; revised on March 22, 2022; editorial decision on April 11, 2022; accepted on April 13, 2022

Abstract

Motivation: Tissue-level omics data such as transcriptomics and epigenomics are an average across diverse cell types. To extract cell-type-specific (CTS) signals, dozens of cellular deconvolution methods have been proposed to infer cell-type fractions from tissue-level data. However, these methods produce vastly different results under various real data settings. Simulation-based benchmarking studies showed no universally best deconvolution approaches. There have been attempts of ensemble methods, but they only aggregate multiple single-cell references or reference-free deconvolution methods.

Results: To achieve a robust estimation of cellular fractions, we proposed EnsDeconv (Ensemble Deconvolution), which adopts CTS robust regression to synthesize the results from 11 single deconvolution methods, 10 reference datasets, 5 marker gene selection procedures, 5 data normalizations and 2 transformations. Unlike most benchmarking studies based on simulations, we compiled four large real datasets of 4937 tissue samples in total with measured cellular fractions and bulk gene expression from different tissues. Comprehensive evaluations demonstrated that EnsDeconv yields more stable, robust and accurate fractions than existing methods. We illustrated that EnsDeconv estimated cellular fractions enable various CTS downstream analyses such as differential fractions associated with clinical variables. We further extended EnsDeconv to analyze bulk DNA methylation data.

Availability and implementation: EnsDeconv is freely available as an R-package from https://github.com/randel/EnsDeconv. The RNA microarray data from the TRAUMA study are available and can be accessed in GEO (GSE36809). The demographic and clinical phenotypes can be shared on reasonable request to the corresponding authors. The RNA-seq data from the EVAPR study cannot be shared publicly due to the privacy of individuals that participated in the clinical research in compliance with the IRB approval at the University of Pittsburgh. The RNA microarray data from the FHS study are available from dbGaP (phs000007.v32.p13). The RNA-seq data from ROS study is downloaded from AD Knowledge Portal.

Contact: jbwang@pitt.edu or wei.chen@chp.edu

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

1 Introduction

The last decades witnessed rich transcriptomics data collected from tissue samples to study complex diseases. However, tissue-level analyses can only obtain the average effects across cell types and are known to be confounded by varying cell-type compositions across samples (Jaffe and Irizarry, 2014). Moreover, the confounding effect is not only additive but also multiplicative (Zheng *et al.*, 2017). Thus it is biologically pressing to assess the impact of cell-type fractions. Technologies such as flow cytometry and immunohistochemistry have been used for cell counting and sorting, but these measurement approaches are too costly to efficiently scalable to large studies. Consequentially, cell counts are usually not measured in tissue samples or hard to measure in solid tissues such as the brain. In most cases, we need to rely on numerical methods, referred to as cell-type deconvolution, to infer cell-type fractions. Fortunately, the growing number of single-cell RNA-sequencing (scRNA-seq) reference datasets has recently galvanized interest in computational approaches to infer cell-type composition.

Cell-type deconvolution methods fall into three broad categories as classified by input: reference-free, partial reference-free and reference-based deconvolution methods (Li and Wu, 2019). The reference-free methods infer cell-type compositions only based on tissue or mixture data itself, and they usually suffer from low accuracy and difficulty interpreting the estimated components (Li and Wu, 2019). The partial reference-free methods require genes that are uniquely expressed in certain cell types, commonly known as cell-type marker genes, to perform deconvolution when the reference is unavailable (Gaujoux and Seoighe, 2013; Zhong *et al.*, 2013). The reference-based methods require reference data from purified cells or scRNA-seq, and most of them focus on cell-type marker genes that can be detected from the reference. They usually utilize a signature matrix [cell-type-specific (CTS) expression averaged from references] to infer cell-type fractions in a regression-based approach. Studies suggest that reference-based methods are the most accurate among the three categories when reliable references are available (Avila Cobos *et al.*, 2020; Hunt *et al.*, 2019). We thus focus on reference-based methods in this study.

To date, there have been several benchmarking studies to assess the performance of cell-type deconvolution methods. For instance, Avila Cobos *et al.* (2020) conducted a simulation-based benchmarking study and highlighted factors that are associated with the performance of cell-type deconvolution, including data normalization, transformation, marker gene selection and choice of deconvolution methods. Nevertheless, unlike real data, the simulated linear mixture may not fully capture biological complexity and noise. More recently, Nadel *et al.* (2021b) compared several deconvolution methods using the Framingham Heart Study (FHS) comprised of thousands of samples with bulk gene expression and cell counts (Mahmood *et al.*, 2014). However, these studies did not demonstrate a universally best computational method to infer cell-type fractions across different tissues (Jin and Liu, 2021). The deconvolution performance depends on various factors in the real data application (Avila Cobos *et al.*, 2020; Nadel *et al.*, 2021b). To resolve some of these issues, SCDC (Dong *et al.*, 2021) proposed to integrate deconvolution results across multiple single-cell references to reduce the impact of potential batch effects of different reference panels. Similarly, DeCompress (Bhattacharya *et al.*, 2021) also studied ensemble deconvolution but limited to reference-free methods.

Here we introduce EnsDeconv, a new ensemble learning-based deconvolution method that comprehensively considers five important factors for cell-type deconvolution, including deconvolution methods, reference datasets, marker gene selection procedures, data normalizations and transformations. We evaluated the impact of the five factors and utilized an ensemble learning approach to integrate the results of various combinations of those factors. We compared our method against existing deconvolution methods in various scenarios using four real bulk datasets with measured cell counts, including three blood datasets and one brain dataset. We define a scenario as a particular setting with a specific reference dataset, marker selection approach, normalization, transformation and deconvolution method. All blood mixtures are large-scale bulk data with hundreds to thousands of samples, capable of providing representative results. The EnsDeconv algorithm is rigorously designed with a CTS robust regression to ensure that our ensemble learning method can provide stable and accurate deconvolution results. Via real-data benchmarking, our method outperforms existing deconvolution approaches in both blood and brain tissues. We also showed that our method can provide helpful information for downstream analyses such as differential cell-type fraction analysis related to clinical variables and extended it to bulk DNA methylation (DNAm) data.

2 Materials and methods

2.1 Review of cellular deconvolution

In general, the cellular deconvolution model can be written as

$$\underset{(G\times S)}{\mathbf{Y}} \approx \underset{(G\times K)}{\mathbf{B}} \times \underset{(K\times S)}{\mathbf{P}},$$

where $\mathbf{Y}$ represents bulk expression for G genes in S samples, $\mathbf{B}$ is the average gene expression over samples for K cell types and $\mathbf{P}$ is the mixing proportions of K cell types per sample, which is usually assumed to be non-negative and a sum of one for each sample. The target of reference-based deconvolution is to estimate cell-type fractions $\mathbf{P}$, with observed bulk expression $\mathbf{Y}$ and average CTS expression $\mathbf{B}$. For each sample s, the cell-type fractions ($\hat{\mathbf{p}}_s$) are estimated via

$$\underset{\mathbf{p}_s\in\Delta_{K-1}}{\operatorname{argmin}} \underbrace{\sum_{g=1}^{G} \ell(y_{g,s}, \mathbf{p}_s^T\mathbf{b}_g)}_{Loss} + \overbrace{\lambda\mathcal{R}(\mathbf{p}_s)}^{Regularizer},$$

where $\Delta_{K-1} = \{x \in \mathbb{R}^K : x_k \geq 0 \text{ and } \sum_{k=1}^{K} x_k = 1\}$, $y_{g,s}$ is the bulk expression for gene g in sample s, $\mathbf{p}_s$ is the mixing proportions of K cell types in sample s, and $\mathbf{b}_g$ is the average CTS gene expression for gene g. The regularizer adds additional constraints to improve the stability of estimated cell-type proportions (Mohammadi *et al.*, 2017).

2.2 The EnsDeconv algorithm

To achieve robust estimation, EnsDeconv considers important factors in the multi-step cellular deconvolution: choice of reference datasets, marker gene selection, normalization and transformation of bulk and reference data and deconvolution methods (Fig. 1A). We exploited all possible combinations of these factors when appropriate and utilized CTS robust regression to obtain the optimal ensemble of cellular fractions, given the true cell-type fractions are often not measured (Fig. 1B).

2.2.1 Single deconvolution methods

To be focused, we only considered peer-reviewed cellular deconvolution methods for gene expression data that estimate cellular fractions not enrichment score, with a customized R package or code, and work for general tissue types. Under those constraints, we investigated the following reference-based deconvolution methods: weighted least squares (LS)—EPIC (Estimating the Proportions of Immune and Cancer cells) (Racle *et al.*, 2017); non-negative LS with normalization—GEDIT (Gene Expression Deconvolution Interactive Tool) (Nadel *et al.*, 2021a); robust regression—FARDEEP (Fast And Robust DEconvolution of Expression Profiles) (Hao *et al.*, 2019); support-vector regression—CIBERSORT (Newman *et al.*, 2015); hybrid scale method—hspe (hybrid-scale proportions estimation) (Hunt and Gagnon-Bartsch, 2021) and dtangle (Hunt *et al.*, 2019); penalized regression with elastic net regularization—DCQ (Digital Cell Quantification) (Altboum *et al.*, 2014); quadratic programming—DeconRNASeq (Gong and

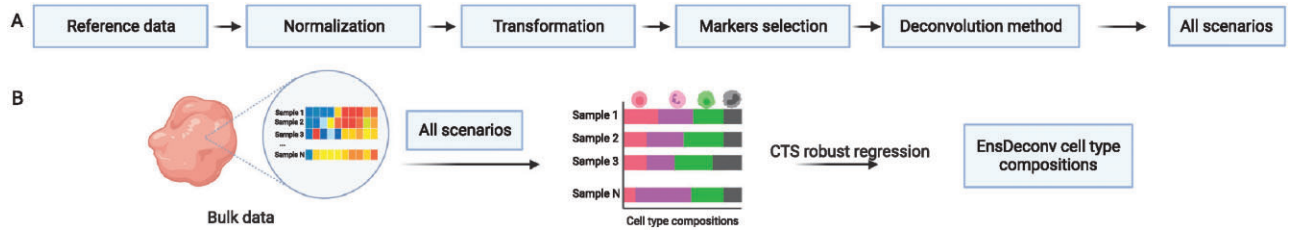

Fig. 1. Overview of the proposed EnsDeconv algorithm. (**A**) Important factors and different deconvolution scenarios (combinations of the five factors). (**B**) Flow chart of EnsDeconv

Szustakowski, 2013); and log-normal model—ICeDT (Immune Cell Deconvolution in Tumor tissues) (Wilson *et al.*, 2020). We also considered two deconvolution methods that are designed for scRNA-seq reference: MuSiC (Multi-subject Single Cell deconvolution) (Wang *et al.*, 2019) and Bisque (Jew *et al.*, 2020). More details about those single deconvolution methods can be found in Supplementary Table S1 and Supplementary Note.

2.2.2 Reference datasets

With the emerging scRNA-seq data, there is a rich collection of published reference datasets we can choose from. For instance, STAB (Song *et al.*, 2021) curated brain scRNA-seq data from 12 studies. This is only a subset of numerous available datasets, and there are cell atlases for different tissues. Recent benchmarking suggested that the best choice of reference data varies between deconvolution methods (Nadel *et al.*, 2021b). To deconvolve a target bulk expression data, we need to check the reference datasets for species, tissue type or brain region, age, disease status, etc. With this consideration, we included seven single-cell reference datasets and a purified-cells dataset to deconvolve brain expression data. The reference data sources vary from different technical platforms, and they are all from cortex tissues of adults (Supplementary Table S2). Three reference datasets are used for blood expression deconvolution, including the widely used lm22 (Newman *et al.*, 2015), skin signatures (Swindell *et al.*, 2013) and ImmunoStates (Vallania *et al.*, 2018). All the three are microarray references since scRNA-seq is hard to detect eosinophil, a rare but key cell type associated with atopy in our application.

2.2.3 Marker gene selection

Most deconvolution algorithms rely on cell-type marker genes (Avila Cobos *et al.*, 2020; Hunt and Gagnon-Bartsch, 2021). While recent studies compared several marker gene selection approaches (Avila Cobos *et al.*, 2020; Hunt *et al.*, 2019), the impact of marker gene selection on deconvolution is unclear. It is generally unknown to choose marker genes that are consistently optimal for all scenarios. We comprehensively considered five marker gene selection approaches (Supplementary Table S3). Based on our study, some marker gene selection approaches may fail to select appropriate markers for certain cell types. If some of the cell-type markers are not available, this specific scenario will be dropped. Since MuSiC utilizes gene weights, all genes will be considered. For ImmunoStates, we used all marker genes from the signature matrix.

Most studies in the field of cell-type deconvolution have only focused on one or several of the marker gene selection approaches. No benchmarking studies appear to have considered the effects of the marker gene selection procedure on results of real data deconvolution. Our work provides additional insights into the performance of different marker gene selection approaches. Some deconvolution methods require a signature matrix (**B** for marker genes only). Given a reference expression matrix **R** for C cells or purified-cell samples, assuming $c \in \mathbf{c}_k$ denoting cell index belonging to the k-th cell type. The element in signature matrix is calculated as $b_{g,k} = \sum_{c \in \mathbf{c}_k} r_{g,c}/|\mathbf{c}_k|$, where $|\mathbf{c}_k|$ is the number of cells in the k-th cell type.

2.2.4 Data normalization and transformation

Different data normalization and transformation are applied to mixture datasets and reference datasets. For RNA-seq data or scRNA-seq data, we considered the following normalization: (i) raw read counts (i.e. no normalization), (ii) count per million (CPM), (iii) transcript per million (TPM) and (iv) trimmed mean of M-values (Robinson and Oshlack, 2010). For deconvolution of blood data, we applied two approaches: (i) keep the data in the original scale or (ii) quantile normalization, since reference data (microarray) and mixture data (RNA-seq) (Jiang *et al.*, 2019) come from different technology. MuSiC and Bisque recommended that both mixture samples and reference data should take CPM normalization. Some other studies also have their recommendations on data normalization, like joint quantile normalization for GEDIT. However, data normalization and transformation are usually not explicitly discussed in most of the deconvolution methods.

After data normalization, the next step is to perform data transformation. Some deconvolution methods provided a recommended transformation approach and some required data on a specific scale. The first question is about taking log transformation or not. In general, a linear scale seems more biologically plausible (Zhong and Liu, 2011), while log-transformation stabilizes the data. Recently, some methods attempted to combine the two scales (Hunt *et al.*, 2019; Hunt and Gagnon-Bartsch, 2021; Wilson *et al.*, 2020). To be consistent, we applied the same data normalization and transformation approach to both bulk and reference datasets. To note, for the unique molecular identifier (UMI)-based 10× reference datasets, we did not apply TPM normalization since UMI counts are invariant to gene length. Therefore, we allowed the normalizations for bulk samples and reference datasets to differ for these reference datasets.

2.2.5 Estimating ensemble cellular fractions

With consideration of all appropriate combinations of the factors mentioned above in deconvolution, we utilized ensemble learning to synthesize $\hat{\boldsymbol{P}}_1, \ldots, \hat{\boldsymbol{P}}_D$, the estimated cellular proportions from each of the D scenarios. Since we typically do not have true proportions, we do not know which estimates to prioritize. Instead, we assume that most estimates resemble true proportions, but some may be outliers. Given the equivalence between outlier detection and robust regression (She and Owen, 2011), this suggests that we can cast this ensemble learning problem as the following robust regression problem:

$$\underset{\substack{\boldsymbol{W}_1,\ldots,\boldsymbol{W}_K \in [0,1]^S \\ (\boldsymbol{W}_1,\ldots,\boldsymbol{W}_K)1_K = 1_S}}{\operatorname{argmin}} \sum_d \sum_{k=1}^{K} \|\hat{\boldsymbol{W}}_{dk} - \boldsymbol{W}_k\|_2, \tag{1}$$

where $\boldsymbol{W}_k$ is the k-th cell type's ensemble fraction, $\hat{\boldsymbol{W}}_{dk}$ is deconvolution scenario d's estimate for the k-th cell-type fraction, and $\|\boldsymbol{v}\|_2 = (\sum_i v_i^2)^{1/2}$ is a vector analogue of absolute deviation. Since the objective in (1) is separable by cell type, (1) allows each cell type to have different outlying deconvolution scenario's. For example, one set of cell-type markers may poorly estimate eosinophil fractions, but may be optimal for monocytes. Since (1) is equivalent to minimizing $\tilde{f}(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{K-1}) = \sum_d \sum_{k=1}^{K-1} \|\hat{\boldsymbol{W}}_{dk} - \boldsymbol{W}_k\|_2 +$

Algorithm 1: EnsDeconv

Notation: α, step size

Input: Estimated fractions of K cell types from D scenarios,

$(\hat{\boldsymbol{W}}_{11},\ldots,\hat{\boldsymbol{W}}_{1K}),\ldots,(\hat{\boldsymbol{W}}_{D1},\ldots,\hat{\boldsymbol{W}}_{DK})$.

Output: EnsDeconv estimated cellular fractions $\hat{\boldsymbol{P}}$.

$t \leftarrow 0$;

$$(\boldsymbol{W}_1^t,\ldots,\boldsymbol{W}_{K-1}^t) \leftarrow \frac{\sum_d(\hat{\boldsymbol{W}}_{d1},\ldots,\hat{\boldsymbol{W}}_{dK-1})}{D};$$

while *not converged* **do**

$$(\boldsymbol{Z}_1^{t+1},\ldots,\boldsymbol{Z}_{K-1}^{t+1}) \leftarrow (\boldsymbol{W}_1^t,\ldots,\boldsymbol{W}_{K-1}^t) - \alpha^t\nabla_{\mathbf{W}_k}\tilde{f}(\boldsymbol{W}_1^t,\ldots,\boldsymbol{W}_{K-1}^t);$$

$$(\boldsymbol{W}_1^{t+1},\ldots,\boldsymbol{W}_K^{t+1}) \leftarrow \mathcal{P}_{\mathcal{C}}(\boldsymbol{Z}_1^{t+1},\ldots,\boldsymbol{Z}_{K-1}^{t+1});$$

$t \leftarrow t+1$;

end

Return $\hat{\boldsymbol{P}} = (\boldsymbol{W}_1,\ldots,\boldsymbol{W}_K)$.

$\sum_d \|\hat{\boldsymbol{W}}_{dK} - (1_S - \sum_{k=1}^{K-1}\boldsymbol{W}_k)\|_2$, we use projected gradient descent to solve this convex problem in Algorithm 1. Note $\mathcal{P}_{\mathcal{C}}(x_0) = \text{argmin}_{x\in\mathcal{C}}\|x - x_0\|_2$ projects x_0 onto $\mathcal{C} = \{\boldsymbol{W}_1,\ldots,\boldsymbol{W}_{K-1} \in [0,1]^S, (\boldsymbol{W}_1,\ldots,\boldsymbol{W}_{K-1})1_{K-1} \leq 1_S\}$.

2.3 Real benchmarking datasets

Instead of relying on simulations that may be oversimplified or unrealistic, we collected and assembled four datasets from both public databases and our local studies. We benchmarked the deconvolution methods with four real mixtures datasets, including one brain dataset and three blood datasets. All four datasets have ground truth of measured cell-type compositions. The Religious Orders Study (ROS) data are from brain dorsolateral prefrontal cortex (DLPFC) tissue of 49 elderly donors with both bulk RNA-seq data (Mostafavi *et al.*, 2018) and measured cell-type fractions (Patrick *et al.*, 2020). The immunohistochemistry-based cell-type proportions are for five cell types: astrocytes (Astro), endothelial (Endo), microglial (Micro), neuron (Neuron) and oligodendrocytes (Oligo).

We used three bulk datasets for blood cells. The first one is Epigenetic Variation and Childhood Asthma in Puerto Ricans (EVAPR). We focused on 220 non-asthma subjects with both bulk RNA-seq expression and measured white blood cell fractions through complete blood count (Jiang *et al.*, 2019). The second dataset, FHS, is composed of two cohorts: the offspring and third-generation cohorts. We downloaded FHS data from dbGaP (phs000007.v32.p13). Here we treated the blood cell counts obtained through a complete blood count using the Coulter HmX Hematology Analyzer, as ground truth. There are 4110 samples from offspring and third-generation cohorts with both cell counts and blood microarray expression. The third dataset (TRAUMA from the Inflammation and Host Response to Injury program) (Xiao *et al.*, 2011) was collected from 167 severe blunt trauma patients under the age of 55 years old (GSE36809). Initially, a blood sample was collected around 12 h of the injury and 1, 4, 7, 14, 21 and 28 days later. There are 558 samples with both white blood cell counts and blood microarray expression. We considered four cell types: neutrophils (Neutro), monocytes (Mono), lymphocytes (Lymph) and eosinophils (Eosino).

We evaluated the accuracy of deconvolution methods using total mean absolute error (MAE) comparing estimated and measured cell-type proportions and calculated the concordance by Spearman's correlation for each cell type. Spearman's correlation is more robust as compared to Pearson's correlation since it is based on rank. Higher mean Spearman's correlation across cell types and lower MAE values indicate a better deconvolution performance. The mean CTS Spearman's correlation is more meaningful than total Spearman's correlation calculated by pooling cell types, which is widely used by most benchmarking studies. The reason is that total correlation is usually overestimated and largely affected by mean fraction values. Instead, the CTS correlation is implicitly used in most downstream analyses to assess sample-level associations.

3 Results

3.1 Benchmarking of factors that impact deconvolution

To understand the importance of those factors we incorporated in EnsDeconv, we ran all possible combinations of those factors to deconvolve the four bulk datasets with measured cell counts. We investigated the overall impact of deconvolution methods, marker gene selection approaches, reference datasets, data normalizations and transformations. We calculated the pairwise correlation of scenarios by varying one factor each time and keeping the other factors unchanged. The lower pairwise correlation indicates higher variability in that factor and thus reflects relatively higher importance (Vallania *et al.*, 2018).

As expected, deconvolution methods play an important role (Supplementary Fig. S1). Detailed comparisons observe that the performance of each single deconvolution method varies greatly across the four bulk datasets (Fig. 2 and Supplementary Fig. S2). Compared with any single deconvolution method, EnsDeconv robustly shows higher Spearman's correlation between estimated and measured fractions across cell types in all benchmarking datasets. The advantage of EnsDeconv is more apparent in rare cell types such as eosinophils in blood.

Reference datasets have the most influential impact on the estimation of cellular fractions (Supplementary Fig. S1), which agrees with previous finding (Vallania *et al.*, 2018). While the reference datasets seem to have consistent performance in three blood cell datasets, their relative performance is generally unknown in solid tissues such as the brain (Supplementary Fig. S3), especially given new reference datasets are being generated. Therefore, it is essential to integrate results from reference datasets. When there are many potential references, we can first filter reference datasets based on sample characteristics such as species, age, and disease status to accelerate the computation time. Ideally, good reference data should be comparable to bulk data in sample characteristics. While many studies have provided useful data sources, for instance, the STAB atlas contains 12 scRNA-seq studies (Song *et al.*, 2021) from the human brain; only 5 datasets share similar age (adult) and brain region (cortex) as ROS bulk data used in this study. For blood data, all the reference data we used are based on microarray, since eosinophils of interest are relatively rare and hard to quantify in scRNA-seq.

The performance of deconvolution also depends on the choice of marker gene selection approach (Supplementary Fig. S3). Through detailed assessment, we selected 50 marker genes per cell type based on benchmarking (Supplementary Fig. S4). We further assessed the data normalization and transformation for each deconvolution method for bulk and scRNA-seq data (Supplementary Figs S3 and S5). Data transformation plays a more important role than data normalization in deconvolution (Supplementary Fig. S1). This finding can guide us in data pre-processing. For deconvolution methods without a clear description of the input data format and those not designed for a specific data scale, it may be beneficial to aggregate the results across different scales (log or linear) to resemble a hybrid scale (Hunt and Gagnon-Bartsch, 2021).

3.2 Ensdeconv improves robustness and accuracy of cellular deconvolution

We have shown that EnsDeconv outperforms all single deconvolution methods in all benchmarking datasets (Fig. 2 and Supplementary Fig. S2). Here for a specific demonstration, we provide an in-depth comparison with the overall best single

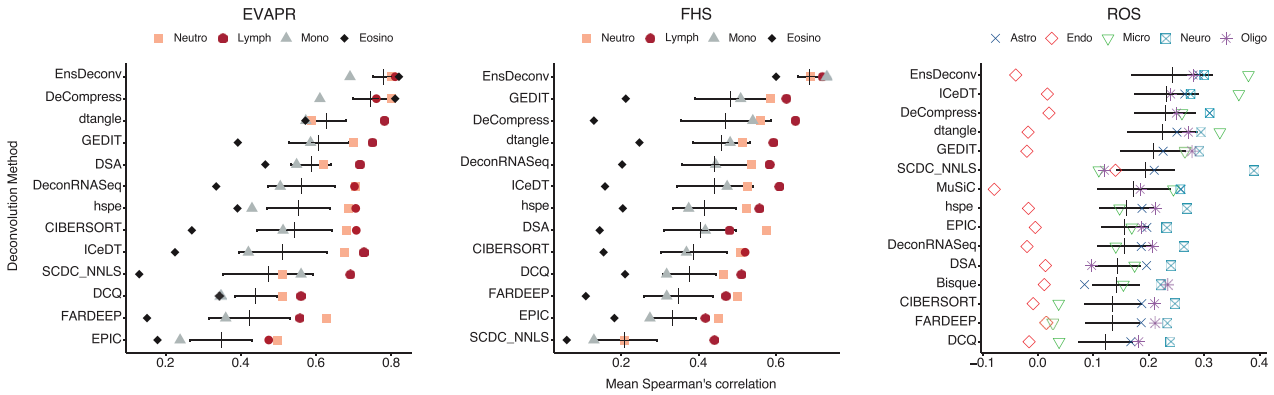

Fig. 2. Comparison of different deconvolution methods and EnsDeconv on EVAPR, FHS and ROS transcriptomics data. Methods designed for scRNA-seq reference (MuSiC and Bisque) are only shown for ROS data since the other three datasets use purified-cells microarray references. The result of TRAUMA data is shown in Supplementary Figure S2. For EnsDeconv, each dot denotes one correlation for each cell type. For other methods, each dot represents the average of Spearman's correlations across scenarios of the particular deconvolution method in each cell type. The black vertical line shows the mean of CTS Spearman's correlations, and the horizontal line presents mean ± standard error of the mean. SCDC_NNLS is the result of integrating all deconvolution scenarios using the NNLS method stated in SCDC. DeCompress is the result of integrating all deconvolution scenarios using the best method that predicts bulk data (Bhattacharya *et al.*, 2021). For completeness, we further compare with a partial reference-free method DSA (Zhong *et al.*, 2013)

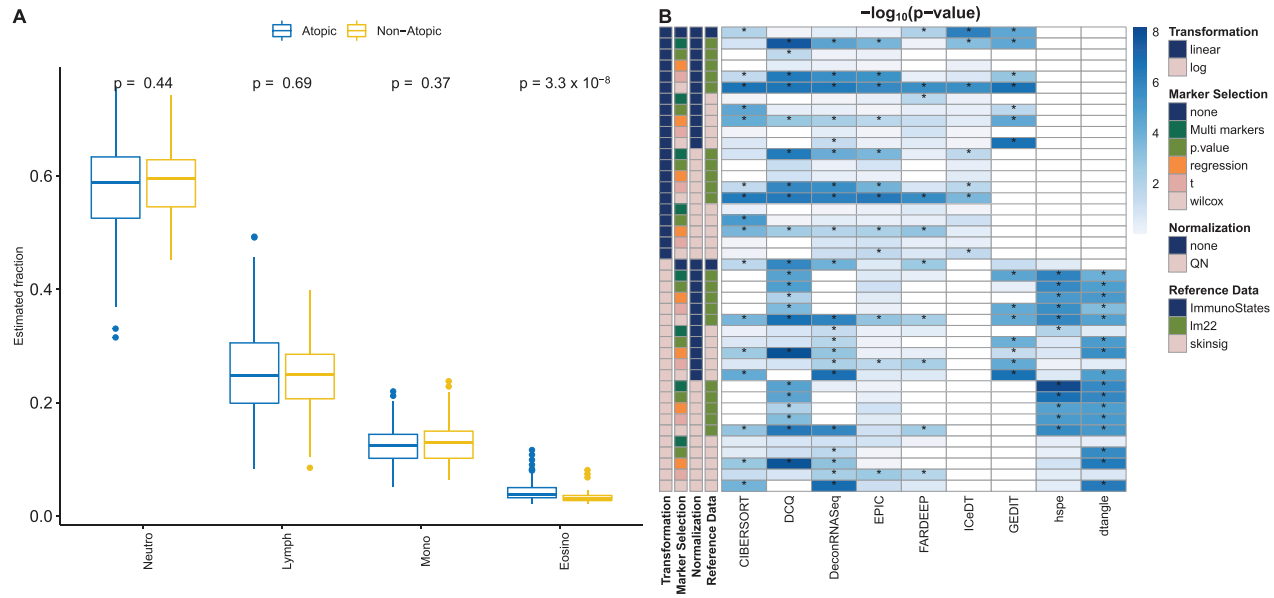

Fig. 3. Analysis of differential cell-type fraction with EVAPR data. (**A**) Comparison using EnsDeconv results for atopic and non-atopic samples. (**B**) Heatmap of −log10(*P*-value) of eosinophils. Each column represents a single deconvolution method. Each row represents a deconvolution scenario combining different factors. For marker selection, 'none' represents using all genes from the ImmunoStates signature matrix (Vallania *et al.*, 2018). QN, quantile normalization. White color denotes specific scenarios that cannot produce results due to model design or computational issues. '*' displayed in the cells indicates scenarios with *P*-value <0.05/4

deconvolution method in our benchmarking, GEDIT (Nadel *et al.*, 2021a). We selected the best scenario for GEDIT based on prediction of bulk data. EnsDeconv leads to more accurate deconvolution results in all datasets with higher correlation concordance than the best scenario of GEDIT (Supplementary Fig. S6). We also compared EnsDeconv with an existing ensemble deconvolution method, SCDC, which is designed to the ensemble of scRNA-seq references only. SCDC shows less accurate results than EnsDeconv, as exampled in the benchmarking of ROS data (Supplementary Fig. S7). This is especially true in the estimation of microglia, a key cell type for Alzheimer's disease. Furthermore, we also compared the ensemble algorithms used in SCDC and DeCompress that utilize bulk data prediction as a surrogate to integrate other factors (Fig. 2 and Supplementary Fig. S2). Since not all deconvolution methods require a signature matrix, we slightly modified the algorithms stated in SCDC and DeCompress. We fitted regression models between the bulk data and predicted cell-type fractions to get predicted bulk data. More details about SCDC and DeCompress' algorithm can be found in Supplementary Note. This result further shows that EnsDeconv can produce more robust result across datasets.

Specifically for brain tissue, we validated EnsDeconv by using it to deconvolve the ROS data in various deconvolution methods and reference datasets (Supplementary Fig. S8). Results show that the mean Spearman's correlation across cell types for EnsDeconv is the best among all deconvolution scenarios (Fig. 2). We further assessed whether EnsDeconv's performance is robust in other tissues by estimating cell fractions of the three blood datasets. We demonstrated that EnsDeconv attains consistently robust results across tissues by observing a similar pattern in blood data as in brain data (Fig. 2 and Supplementary Fig. S2).

3.3 Ensdeconv provides biologically meaningful results for differential fraction analysis

To demonstrate the usage of our method in downstream analyses, we considered the EVAPR dataset, which has a clinical indicator

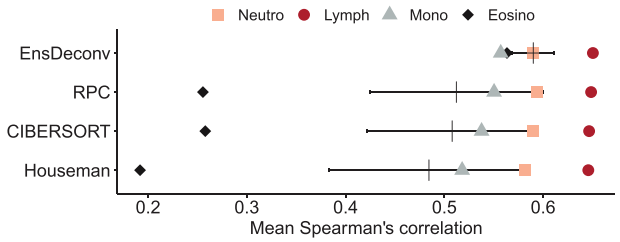

Fig. 4. Comparison of three single deconvolution methods and EnsDeconv on FHS DNAm data. For EnsDeconv, each dot denotes one correlation for each cell type. For other methods, each dot represents the average of Spearman's correlations across scenarios of the particular deconvolution method in each cell type. The black vertical line shows the mean of CTS Spearman's correlations. The horizontal line presents mean ± standard error of the mean

variable for atopy. It is of scientific interest to compare the cell-type fractions between atopic and non-atopic individuals. The comparisons on measured cell-type fractions detect a significant difference only in eosinophils, a type of white blood cells known to be elevated in atopic subjects (two-sided Wilcoxon test P-value $= 1.7 \times 10^{-10}$). With EnsDeconv estimated cell-type proportion, we replicated the significant finding in eosinophils (two-sided Wilcoxon test P-value $= 3.3 \times 10^{-8}$), while there is no significant difference in the other cell types (Fig. 3A).

We repeated the analyses with single deconvolution methods using microarray references in eosinophils under different data normalizations and transformations, marker gene selection approaches, and reference datasets (Fig. 3B). Most deconvolution methods, like dtangle (Hunt *et al.*, 2019), GEDIT (Nadel *et al.*, 2021a), DCQ (Altboum *et al.*, 2014) and DeconRNASeq (Gong and Szustakowski, 2013), can detect the differential fraction between atopic and non-atopic samples under some scenarios but failed in the other cases. In addition, methods like FARDEEP (Hao *et al.*, 2019), CIBERSORT (Newman *et al.*, 2015), and EPIC (Racle *et al.*, 2017) have difficulties discovering differences between atopic and non-atopic samples in eosinophils under most scenarios. We also provided results for the other three cell types (Supplementary Fig. S9). In some cases, some deconvolution methods will detect significant differences with cell types that do not show association in measured cell counts, rendering the results unreliable. This analysis also establishes that none of the reference datasets, marker gene selection approaches, or data normalization and transformation approaches is advantageous in detecting differential cell-type compositions.

3.4 Extension to bulk DNAm data

In addition to gene expression data, FHS also quantified DNAm for a subset of individuals and we deconvolved 3013 blood samples with DNAm beta values and measured cell counts. We used two blood DNAm references (Reinius *et al.*, 2012; Salas *et al.*, 2022). Following the pipeline in the minfi R package (Aryee *et al.*, 2014), we constructed the signature matrices from the two references and also included the signature matrix derived from Reinius *et al.* (2012) in the EpiDISH R package (Teschendorff *et al.*, 2017). With the three signature matrices, we utilized three deconvolution methods: quadratic programming (QP) (Houseman *et al.*, 2012), CIBERSORT (Newman *et al.*, 2015) and robust partial correlations (Teschendorff *et al.*, 2017). After deconvolution, cell subtypes are aggregated to the four major cell types with average measured fraction >1%.

With EnsDeconv on those results from three deconvolution methods and three signature matrices, we observed more robust and accurate results than other methods (Fig. 4). The relatively rare cell-type eosinophil has worse performance with the reference from Reinius *et al.* (2012), but EnsDeconv is less affected by that and can achieve better concordance between measured and estimated cellular fractions.

4 Conclusion and discussion

In summary, we developed an ensemble deconvolution approach (EnsDeconv) to incorporate multiple factors that affect the estimation of cell-type fractions: deconvolution methods, reference datasets, marker gene selection, data normalization and transformation in the estimation of cell-type fractions in large bulk datasets. We benchmarked EnsDeconv with three large real datasets of blood cells and one brain dataset with both bulk gene expression and measured cell counts. Benchmarking shows that EnsDeconv provides more robust and accurate estimates of cell-type fractions than a single deconvolution method across tissues. Compared with the best single deconvolution method identified in the benchmarking, EnsDeconv shows better sample concordance and higher accuracy. For EVAPR data, measured cell counts identify a significant difference between atopic and non-atopic samples only in eosinophil, and EnsDeconv replicates this finding exactly. However, a practitioner may randomly pick a convenient existing deconvolution method with an arbitrary reference dataset, marker selection procedure, data normalization and transformation. The results show dramatic random variability, while EnsDeconv can robustly and accurately integrate those factors.

Previous work mainly focused on either small-scale real data or simulated data that may be unrealistic. We instead compiled large-scale real bulk mixture blood and brain datasets with measured cellular fractions. This study also provides the first investigation on the deconvolution performance of two large datasets, EVAPR and TRAUMA, and also the DNAm data from FHS. Our benchmark analysis systematically investigated how different factors affect deconvolution performance and observe inconsistent results of the existing deconvolution methods that depend highly on references, data transformations and marker genes. In addition to benchmarking, we further utilized ensemble learning of various deconvolution methods to improve deconvolution accuracy and concordance. EnsDeconv downweights potentially worse deconvolution scenarios and takes advantage of each single deconvolution method, reference dataset and marker selection procedure. Finally, our software provides a convenient tool for the users to apply any preferred methods with benchmarking datasets.

This article has some limitations. EnsDeconv goes through various deconvolution approaches under different scenarios to provide a complete picture, so it cannot be as fast as some existing deconvolution methods. By making the computation highly parallel and using 60 computational nodes, we can finish running all scenarios (around 300) for the FHS dataset (4110 samples) in about 5 h. For a smaller sample size, we can finish running the computation in 2–4 h (EVAPR: around 300 scenarios and 220 samples; TRAUMA: about 300 scenarios and 558 samples; and ROS: around 1400 scenarios and 49 samples). Details about the number of scenarios can be found in Supplementary Table S4. Most methods we compared in this study can finish deconvolution in a relatively short time. However, focusing on reference-based methods, hybrid-scale methods such as ICeDT and hspe may need longer running time (Supplementary Fig. S10), so we can adaptively limit the deconvolution scenarios on them. Nonetheless, parallel computing can help us shorten the computation time, and it has been built into our software. Lastly, the sample size for our brain benchmarking dataset is relatively small and may be less representative than blood datasets.

Robust and accurate estimation of cell-type fractions can serve as input for many downstream analyses at cell-type resolution. Representative analyses include CTS differential expression (Wang *et al.*, 2021) and CTS expression quantitative trait loci (eQTLs) (Wang *et al.*, 2020, 2021). Other than working on one random estimate of cell-type fractions, our ensemble algorithm constructs a sensitivity analysis framework to incorporate most top published deconvolution approaches. The varying deconvolution factors provide a complete picture of the confounding caused by cell-type heterogeneity in tissue-level analyses. The framework presented in this work can be easily extended to deconvolving omics data types other than transcriptomics and DNA methylomics and incorporating other deconvolution methods.

Acknowledgements

The results published here are in part based on data obtained from the AD Knowledge Portal. Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161 (ROS), R01AG15819 (ROSMAP; genomics and RNAseq), R01AG17917 (MAP), R01AG30146, R01AG36836 (RNAseq), RF1AG57473 (single nucleus RNAseq), U01AG46152 (ROSMAP AMP-AD, targeted proteomics), U01AG61356 (whole-genome sequencing, targeted proteomics, ROSMAP AMP-AD) and the Illinois Department of Public Health (ROSMAP). The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195, HHSN268201500001I and 75N92019D00031). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Additional funding for SABRe was provided by Division of Intramural Research, NHLBI, and Center for Population Studies, NHLBI.

Funding

This research was funded in part through a grant from the University of Pittsburgh Brain Institute, University of Pittsburgh Medical Center Competitive Medical Research Fund, National Institutes of Health's UL1TR001857, R01HL117191, R21HL150431, R37MH057881, R37MH057881-22S1 and R01MH123184. This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided.

Conflict of Interest: none declared.

References

Altboum,Z. *et al.* (2014) Digital cell quantification identifies global immune cell dynamics during influenza infection. *Mol. Syst. Biol.*, **10**, 720.

Aryee,M.J. *et al.* (2014) Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.

Avila Cobos,F. *et al.* (2020) Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.*, **11**, 5650.

Bhattacharya,A. *et al.* (2021) Decompress: tissue compartment deconvolution of targeted mRNA expression panels using compressed sensing. *Nucleic Acids Res.*, **49**, e48.

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

Gaujoux,R. and Seoighe,C. (2013) Cellmix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics*, **29**, 2211–2212.

Gong,T. and Szustakowski,J.D. (2013) DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-seq data. *Bioinformatics*, **29**, 1083–1085.

Hao,Y. *et al.* (2019) Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. *PLoS Comput. Biol.*, **15**, e1006976.

Houseman,E.A. *et al.* (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics.*, **13**, 86–16.

Hunt,G.J. *et al.* (2019) Dtangle: accurate and robust cell type deconvolution. *Bioinformatics*, **35**, 2093–2099.

Hunt,G.J. and Gagnon-Bartsch,J.A. (2021) The role of scale in the estimation of cell-type proportions. *Ann. Appl. Stat.*, **15**, 270–286.

Jaffe,A.E. and Irizarry,R.A. (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.*, **15**, R31.

Jew,B. *et al.* (2020) Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat. Commun.*, **11**, 1971.

Jiang,Y. *et al.* (2019) Transcriptomics of atopy and atopic asthma in white blood cells from children and adolescents. *Eur. Respir. J.*, **53**, 1900102.

Jin,H. and Liu,Z. (2021) A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biol.*, **22**, 1–23.

Li,Z. and Wu,H. (2019) Toast: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome Biol.*, **20**, 1–17.

Mahmood,S.S. *et al.* (2014) The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet*, **383**, 999–1008.

Mohammadi,S. *et al.* (2017) A critical survey of deconvolution methods for separating cell types in complex tissues. *Proc. IEEE*, **105**, 340–366.

Mostafavi,S. *et al.* (2018) A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. *Nat. Neurosci.*, **21**, 811–819.

Nadel,B.B. *et al.* (2021a) The gene expression deconvolution interactive tool (GEDIT): accurate cell type quantification from gene expression data. *GigaScience*, **10**, giab002.

Nadel,B.B. *et al.* (2021b) Systematic evaluation of transcriptomics-based deconvolution methods and references using thousands of clinical samples. *Brief. Bioinform.*, **22**, bbab265.

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

Patrick,E. *et al.* (2020) Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLOS Comput. Biol.*, **16**, e1008120.

Racle,J. *et al.* (2017) Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *elife*, **6**, e26476.

Reinius,L.E. *et al.* (2012) Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS One*, **7**, e41361.

Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25–R29.

Salas,L.A. *et al.* (2022) Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. *Nat. Commun.*, **13**, 761.

She,Y. and Owen,A.B. (2011) Outlier detection using nonconvex penalized regression. *J. Am. Stat. Assoc.*, **106**, 626–639.

Song,L. *et al.* (2021) STAB: a spatio-temporal cell atlas of the human brain. *Nucleic Acids Res.*, **49**, D1029–D1037.

Swindell,W.R. *et al.* (2013) Dissecting the psoriasis transcriptome: inflammatory-and cytokine-driven gene expression in lesions from 163 patients. *BMC Genomics*, **14**, 527–520.

Teschendorff,A.E. *et al.* (2017) A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. *BMC Bioinformatics*, **18**, 1–14.

Vallania,F. *et al.* (2018) Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. *Nat. Commun.*, **9**, 1–8.

Wang,J. *et al.* (2020) Using multiple measurements of tissue to estimate subject-and cell-type-specific gene expression. *Bioinformatics*, **36**, 782–788.

Wang,J. *et al.* (2021) Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. *Genome Res.*, **31**, 1807–1818.

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

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.

Xiao,W. *et al.* (2011) A genomic storm in critically injured humans. *J. Exp. Med.*, **208**, 2581–2590.

Zheng,X. *et al.* (2017) Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biol.*, **18**, 17.

Zhong,Y. and Liu,Z. (2011) Gene expression deconvolution in linear space. *Nat. Methods*, **9**, 8–9; author reply 9.

Zhong,Y. *et al.* (2013) Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics*, **14**, 89–10.