Genome Biology

METHODOLOGY

Open Access

BLEND: probabilistic cellular deconvolution with individualized single-cell reference integration

Penghui Huang[1*], Manqi Cai[1], Chris McKennan[2] and Jiebiao Wang[1*]

*Correspondence: huangpenghui@pitt.edu; jbwang@pitt.edu

[1] Department of Biostatistics and Health Data Science, University of Pittsburgh, De Soto St, Pittsburgh 15261, PA, USA
[2] Department of Statistics, University of Pittsburgh, S Bouquet St, Pittsburgh 15213, PA, USA

Abstract

Cellular deconvolution estimates cell-type fractions from bulk transcriptomic data, but current methods often overlook cell type-specific expression varying across samples, discrepancies between bulk and single-cell data, or lack guidance on reference data selection and integration. Therefore, we present BLEND, a hierarchical Bayesian method that leverages multiple single-cell reference datasets to perform cellular deconvolution. BLEND estimates cellular fractions accurately by learning the most suitable reference for each bulk sample, accounting for the aforementioned issues. BLEND outperforms state-of-the-art methods in comprehensive benchmarking studies using human brain cortex data and provides reliable insights into Alzheimer's disease progression.

Keywords: Cellular deconvolution, Single-cell RNA sequencing, Bayesian estimation, Gibbs sampling, EM algorithm, Maximum a posteriori estimation

Background

Bulk RNA sequencing (RNA-seq) technology provides a valuable and cost-effective approach to studying gene expression patterns across different tissues or conditions [1]. However, bulk RNA-seq is typically performed on tissues composed of many cell types, where estimating the proportion of each cell type in each sample can facilitate cell type-specific (CTS) analyses and address the confounding caused by cellular heterogeneity [2, 3].

Many in silico cellular deconvolution methods have been developed in recent years to estimate cellular fractions as a computational alternative. Cellular fraction estimates allow analysts to infer how cell type proportions vary with different phenotypes, account for cell type confounding, and facilitate downstream CTS analyses, such as estimating CTS gene expression, CTS differential expression, and CTS expression quantitative trait loci analyses [4, 5].

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Cellular deconvolution models bulk gene expression as the weighted average of CTS gene expression, where the weights are cell type proportions. Among existing methods to estimate cell proportions, reference-based methods have shown more robust performance compared to unsupervised reference-free methods [6, 7]. Reference-based methods rely on single-cell or sorted-cell data to construct a CTS expression reference matrix whose entries indicate each gene's expected expression in each cell type. Due to the challenges of incorporating inter-subject variation in CTS expression into deconvolution models, existing methods assume the same reference matrix shared across all study subjects and provide no guidance to choose appropriate references. This belies the fact that there is often substantial heterogeneity in CTS expression that arises due to differences in experimental conditions, technical batch effects, and variation in reference sequencing technologies. Nonetheless, the choice of reference is the most important factor affecting deconvolution accuracy [7–9]. It is therefore critical to develop new statistical methods to individualize signature matrices for each subject to account for inter-subject variation in CTS expression.

Furthermore, the deconvolution of certain tissue types is more challenging due to inherent data characteristics. For example, human brain tissues are typically collected from frozen samples with relatively small sample sizes, and only nuclear RNAs are quantified, leading to single-nucleus RNA sequencing references of generally lower quality compared to other tissues [10]. Also, Sutton et al. [11] highlighted high transcriptomic heterogeneity of human brain due to diverse alternative splicing isoforms, non-coding RNAs, subtypes, etc. These characteristics underscore that the human brain is a highly heterogeneous tissue type with limited quality references.

To address the above issues, we developed BLEND, a hierarchical Bayesian model that is, to our knowledge, the first deconvolution method to leverage multiple available references to create individualized references for study subjects. Inspired by Latent Dirichlet Allocation (LDA) [12], we use a "bag-of-words" representation for bulk RNA-seq count data whereby we view a read count as a word, a cell type as a topic, and a bulk sample as a document. In this analogy, a cell-type topic's reference is its distribution over words. However, unlike conventional LDA-based deconvolution methods, which assume references are shared across bulk samples, BLEND allows references to be sample-specific and uses the data to learn each sample's most appropriate reference among all possible references in the convex hull of available references. We show through extensive realistic simulations and real data applications with a focus on human brain cortex that such reference learning drastically improves cell type proportion estimates. BLEND also stands out for being user-friendly, as it avoids the need for manual reference evaluation, data transformation and normalization, and the selection of cell type marker genes.

Results

Overview of the BLEND method

BLEND is a cellular deconvolution method that can integrate multiple references as visualized in Fig. 1. First, BLEND individualizes the most suitable reference matrix for each bulk sample by exploring convex hulls of available CTS reference vectors. In our context, a convex hull is the set of all possible convex combinations of available cell type reference profiles. Second, BLEND uses individualized references to deconvolve bulk

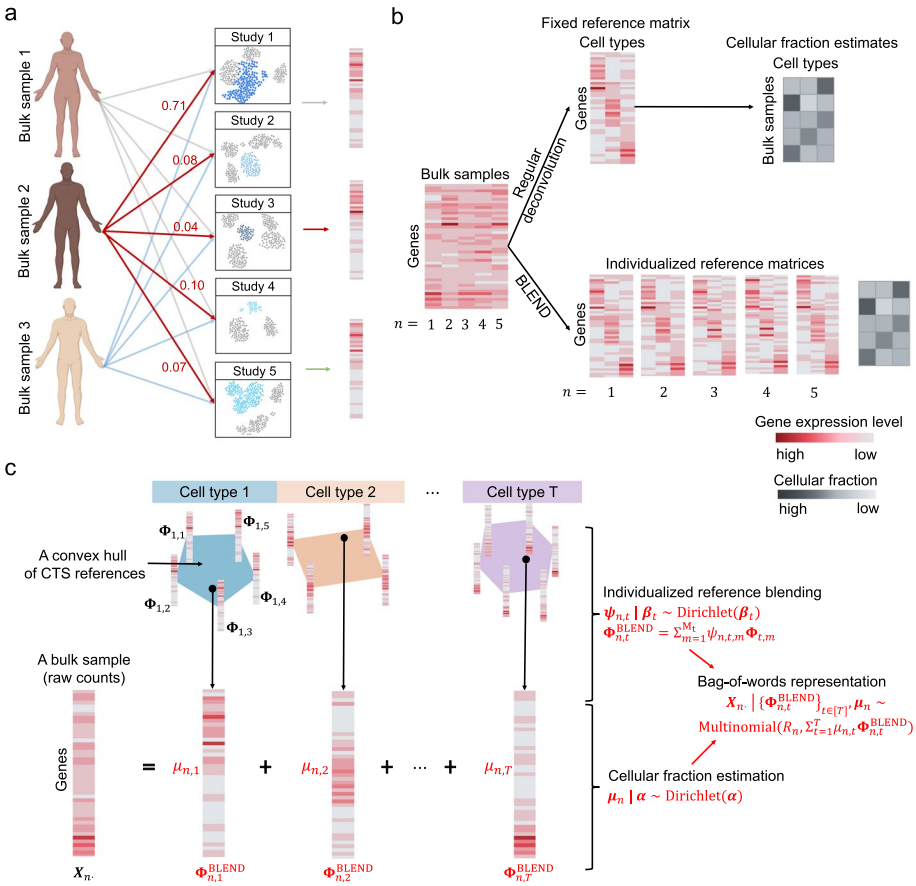

Fig. 1 Overview of BLEND. **a** BLEND integrates multiple cell type-specific (CTS) expression references from single-cell or sorted-cell studies to individualize references for heterogeneous bulk RNA-seq samples. By exploring the convex hull of available references, BLEND automatically selects the most suitable reference for each sample through reference blending. **b** BLEND estimates each bulk sample's cellular fractions using its individualized reference matrix. By contrast, regular deconvolution methods use a fixed reference for all bulk samples. **c** For each bulk sample, BLEND models gene expression as a weighted average of CTS expression, where the weights correspond to cell type fractions. By employing a "bag-of-words" representation, BLEND jointly models individualized references, cellular fractions, and bulk samples in a probabilistic framework

count data by employing a "bag-of-words" representation. BLEND adopts a unified hierarchical Bayesian framework to incorporate these two steps. Two consistent parameter estimation strategies are provided: a Gibbs sampler and an expectation maximization algorithm to maximize the posterior distribution (EM-MAP). We conducted experiments using the Gibbs sampler because of its guaranteed posterior convergence. The detailed statistical model of BLEND is provided in Methods and algorithm derivations are in Additional file 1: Section 1.

BLEND improves performance with partially matched references

In the BLEND model, we assume the CTS expression of bulk samples lies in the convex hulls of the provided references. We first tested BLEND with an oracle simulation that simulated data according to BLEND's generative model, which confirmed that BLEND accurately selects references and estimates cellular fractions (Additional file 1: Fig. S3). However, since real data likely deviate from this ideal setting, we consider more realistic

simulated data below that use real single nucleus RNA-seq (snRNA-seq) data to generate pseudo-bulk data. Notably, these simulations make no assumptions on the distribution of gene counts and do not require a subject's references exactly lie in the convex hulls of provided references.

The real snRNA-seq dataset we used for simulation is from Mathys et al. [13], who collected 427 individuals' postmortem dorsolateral prefrontal cortex (DLPFC) tissues from Religious Orders Study and Rush Memory and Aging Project (ROSMAP) [14]. They measured gene expression levels across 2.3 million nuclei isolated from these tissues using droplet-based snRNA-seq. Among all donors, 418 of them contained cells of six major cell types in the brain: astrocyte (astro), microglia (immune), inhibitory neuron (inh), oligodendrocyte progenitor cell (OPC), oligodendrocyte (oligo), and excitatory neuron (ex). As the data we used were already processed by the original authors to filter out low-quality cells and doublets [13], we did not perform further filtering to keep the data as raw as possible to avoid any biases (e.g., "cherry-picking" samples). To generate realistic data, we randomly chose 40 out of 418 individuals (9.6%) to serve as references to deconvolve all 418 individuals. We generated pseudo-bulk samples by summing up counts of snRNA-seq data for each individual and used the cell counts of different cell types to calculate ground truth cellular fractions.

We compared BLEND with four deconvolution methods: BayesPrism [15], which is an LDA-based Bayesian method using a fixed reference; MuSiC [16], a weighted non-negative least squares method that utilizes multi-subject references; Bisque [17] and hspe [18], both of which were recommended for brain tissue deconvolution by a recent benchmarking paper [19]. We set function parameters as recommended by the original method papers and software vignettes (Additional file 1: Table S2). BLEND used the references from the abovementioned 40 subjects to deconvolve all 418 pseudo-bulk datasets. For each cell type, we calculated Lin's concordance correlation coefficient (CCC) [20] between the true and estimated fractions across individuals. CCC ranges between minus one and one and measures the deviation from the line $y = x$, where values close to one indicate points lie closer to the line. CCC combines the mean difference and correlation between two sets of points to capture both location and scale shifts from the ground truth. We also computed the mean CCC across cell types (mCCC).

We first checked the results for the 40 subjects with matched snRNA-seq reference and pseudo-bulk data. Notably, BLEND was a priori unaware of the correspondence between references and bulk samples and had to learn it from the data. Figure 2a gives the CCC computed for these 40 subjects and shows that BLEND's estimates are superior to those from existing methods. Evaluating the estimated reference mixing proportions of the pseudo-bulk samples with matched references, we found that BLEND was also able to identify matched references for all subjects (Fig. 2b).

We next assessed the results for the remaining 378 pseudo-bulk samples without matched references. Despite not having matched references, we expected BLEND to outperform existing methods by identifying references for each subject that were similar to their CTS expression. Figure 2c plots each method's estimated fractions against the simulated ground truth proportions. BLEND clearly has the largest CCC for all cell types, indicating BLEND's individualized references substantially improve cell type proportion estimates.

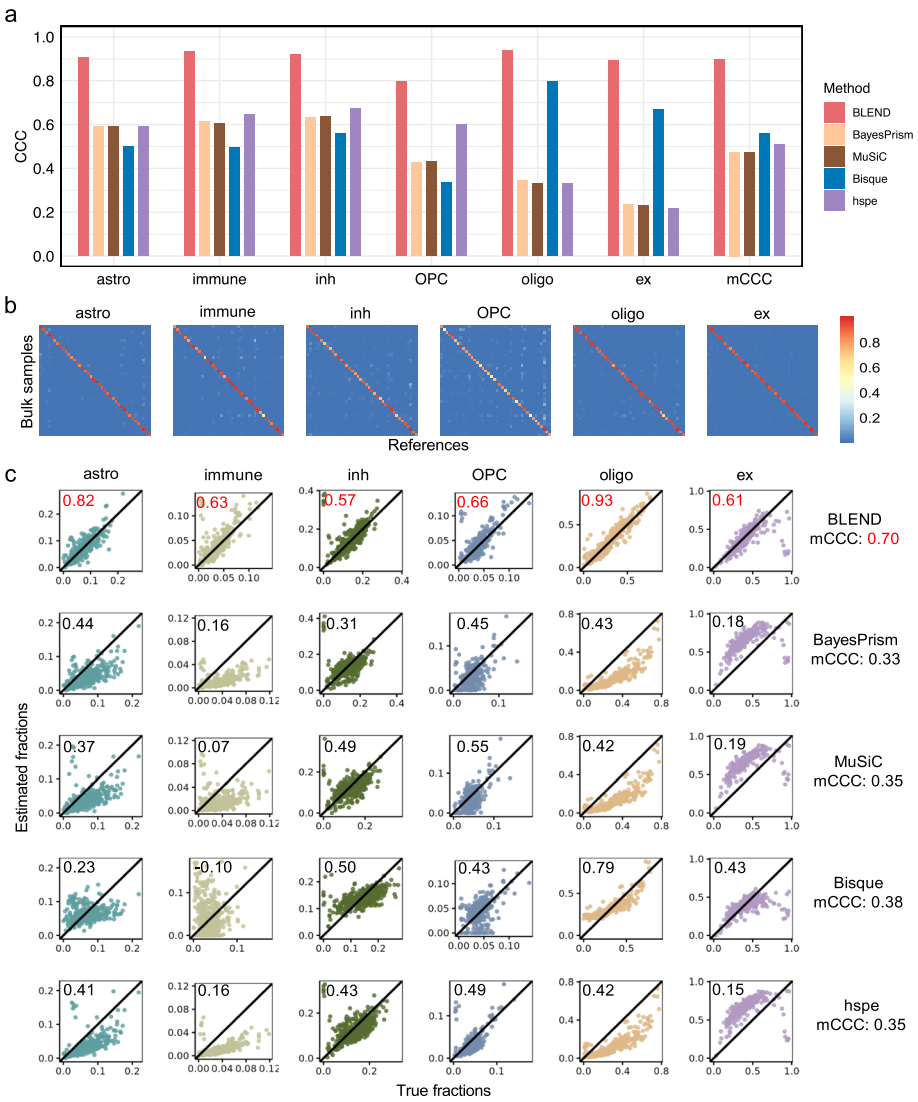

Fig. 2 Benchmarking cellular deconvolution methods with partial reference simulation. **a** Barchart of Lin's concordance correlation coefficients (CCC) between the estimated and ground truth fractions for each cell type across 40 pseudo-bulk samples with matched references and mean CCC (mCCC) across cell types. **b** Heatmaps of estimated reference mixing proportions of the pseudo-bulk samples with matched references. Rows represent bulk samples, and columns represent references. **c** Scatter plots of 378 samples with no matched reference. The x-axis represents true cellular fractions, and the y-axis corresponds to the estimated ones. CCC calculated for each cell type between estimated and true fractions is presented on the top-left corner of each plot, and mCCC across cell types of each method is listed on the right. The CCC of the best performer for each cell type is highlighted in red. astro: astrocyte; inh: inhibitory neuron; OPC: oligodendrocyte progenitor cell; oligo: oligodendrocyte; ex: excitatory neuron

BLEND alleviates discrepancies between bulk and reference data

In most applications, the bulk data to be deconvolved do not have matched reference data, where bulk and reference datasets typically derive from different labs and undergo different processing steps. To examine the impact of such heterogeneity on deconvolution estimates and whether BLEND can circumvent these issues, we considered snRNA-seq data collected on the same set of 231 subjects from Mathys et al. [13] and Fujita et al. [21]. Despite being collected on the same subjects, the CTS expression in these

two datasets shows significant batch effects, where a clear decision boundary separating these studies exists in the uniform manifold approximation and projection (UMAP) plot in Fig. 3a. As such, we used these datasets to illustrate how BLEND overcomes discrepancies between bulk and reference data.

We designed two sets of cross-data simulation studies (Fig. 3b). The first set, denoted "Mathys-Fujita," used the Mathys et al. snRNA-seq data as references to deconvolve pseudo-bulk data generated by collapsing the Fujita et al. snRNA-seq data. The second set, denoted "Fujita-Mathys," employed the opposite approach. In both settings, we generated pseudo-bulk samples and acquired true cell fractions as described in the above-mentioned partial reference simulation. After deconvolution, we calculated the CCC between estimated and ground truth fractions for each cell type across all pseudo-bulk samples and mean CCC across cell types in each simulation setting (Fig. 3c and d). In the first setting, BLEND showed superior performance in all cell types, with a mean CCC of 0.41 higher than the runner-up (Bisque). In the second setting, BLEND performed the best in five out of six cell types and had the highest mean CCC.

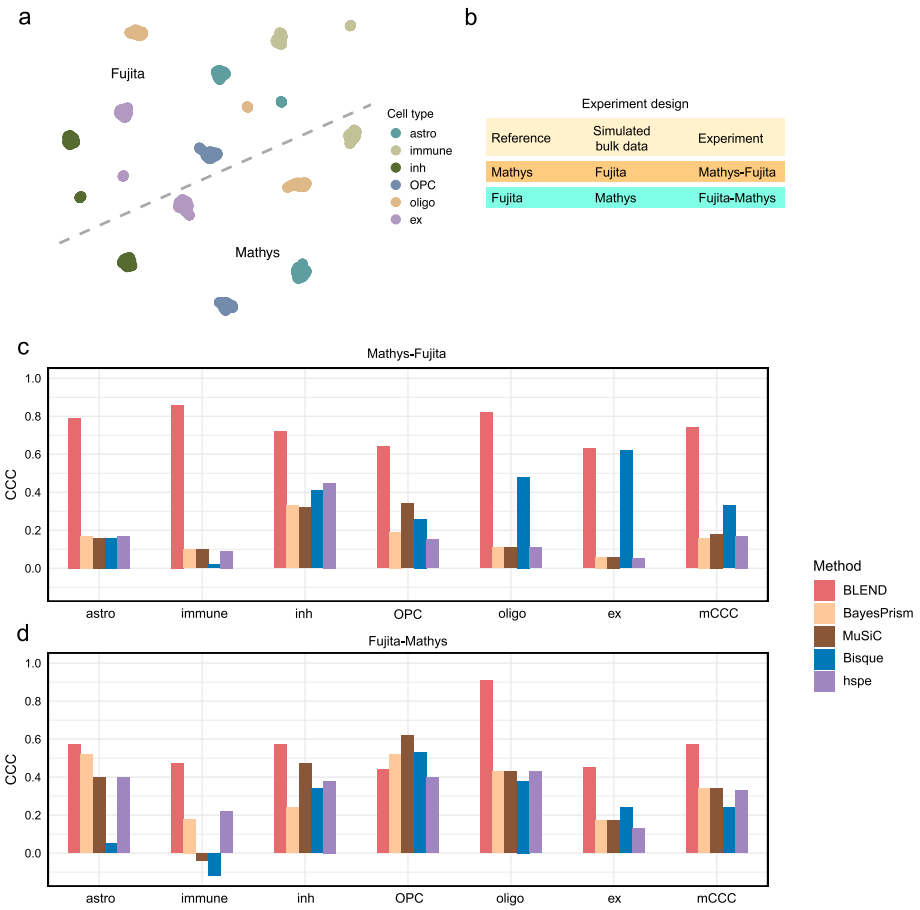

Fig. 3 Benchmarking cellular deconvolution methods with cross-data simulation. **a** UMAP of CTS expression for the 231 individuals from Mathys et al. and Fujita et al.. Each point is a subject, cell type, and experiment triplet. **b** Cross-data simulation experimental design. **c**, **d** The CCC between estimated and ground truth fractions for each cell type across all bulk samples and mCCC across cell types in the cross-data simulation. astro: astrocyte; inh: inhibitory neuron; OPC: oligodendrocyte progenitor cell; oligo: oligodendrocyte; ex: excitatory neuron

BLEND performs well when cell types are missing

Cell type missingness is a common challenge in cellular deconvolution analyses and can occur either in bulk samples (zero fraction) or reference datasets (novel cell type with unknown reference). To systematically evaluate BLEND's robustness under these two scenarios, we conducted experiments addressing each case separately.

First, we considered missingness in bulk data, where a particular cell type may be absent due to sample collection or processing issues. We repeated the experiment described in the partial reference simulation, generating 418 pseudo-bulk samples but intentionally excluding immune cells, which comprise a median of 3.5% of the total cell population. BLEND was provided with reference data for all six cell types, including immune cells. In this scenario, BLEND correctly estimated negligible fractions for the missing immune cells (mean estimated fraction of 1.5×10^{-5} for samples with matched references and 3.7×10^{-3} for samples with no matched reference), without compromising the accuracy for the other five cell types, as demonstrated by barcharts in Fig. 4a.

Next, we investigated the case where the reference data lacks information about a novel cell type that might be present in bulk samples. Since novel or unknown cell types may be encountered in practice, we recommend augmenting the reference dataset with a uniformly distributed surrogate reference across genes, similar to the approach by [22]. To validate this strategy, we again used the 418 pseudo-bulk samples (including immune cells), provided references in which actual immune cell references were replaced by the surrogate reference. Results indicated that BLEND maintained accurate fraction estimates for the five known cell types; however, estimates for the immune cells (the novel cell type without a direct reference) were less precise (Fig. 4b), as expected due to the surrogate reference's inherent deviation from the actual cell type gene expression distribution.

BLEND accurately estimates cellular fractions in real data

Having shown BLEND significantly outperforms existing methods in pseudo-bulk data, we next assessed its performance in real bulk data. To evaluate BLEND in as wide a variety of datasets as possible, we considered data from Huuki-Myers et al. [19], which provided a multi-assay dataset to benchmark deconvolution methods. Their data consisted of postmortem human dorsolateral prefrontal cortex (DLPFC) samples from 10 donors and were assayed from adjacent brain regions using three RNA extraction methods (nuclear [Nuc], cytoplasmic [Cyto], and total [Bulk]) and two library preparation techniques (polyA and RiboZeroGold). To benchmark deconvolution methods, Huuki-Myers et al. [19] also measured the fractions of astrocytes, microglia, oligodendrocytes, endothelial cells, inhibitory neurons, and excitatory neurons via RNAScope/Immunofluorescence (RNAScope/IF). A total of 110 bulk samples with RNAScope/IF cell type proportions passed the quality control and were provided in the dataset. They also included 19 snRNA-seq reference samples measured from the same 10 donors used to generate bulk samples. Following Huuki-Myers et al. [19], we used the 19 snRNA-seq samples as references.

We evaluated each method's performance on these data by computing Pearson's correlation coefficient between estimated and RNAScope/IF cellular fractions. We

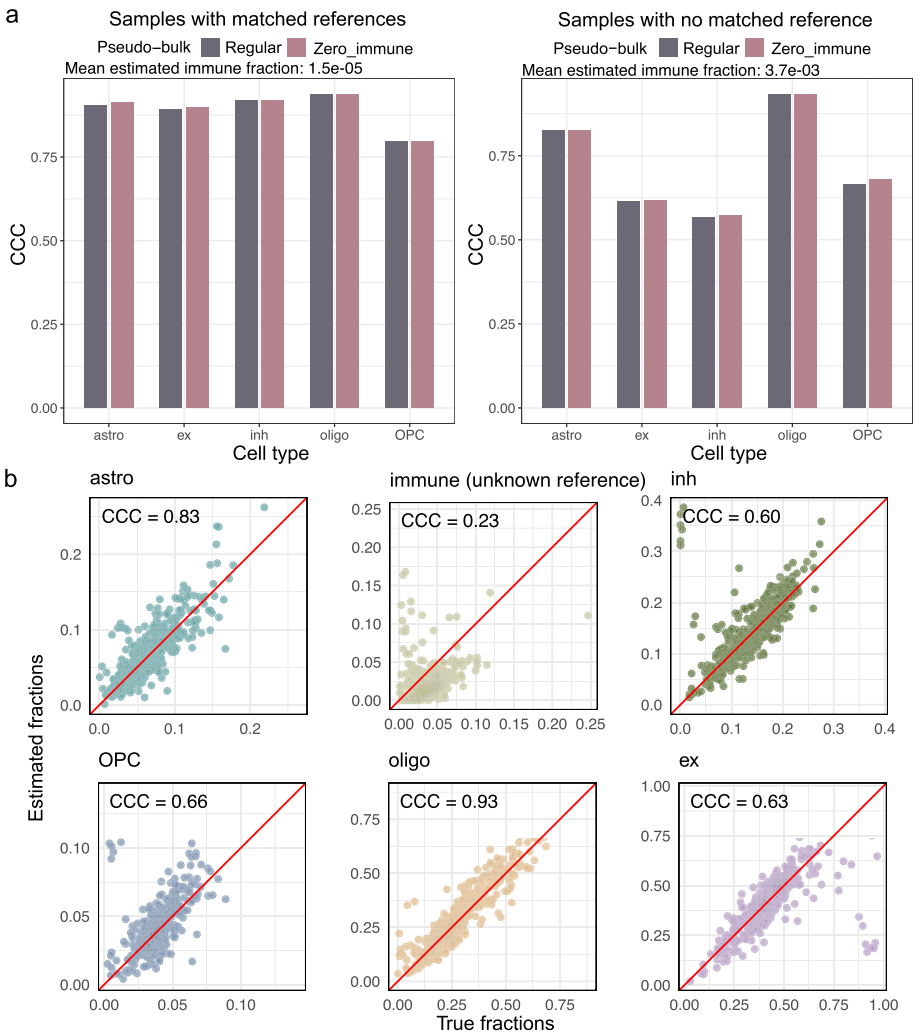

Fig. 4 BLEND's performance under cell type missingness in bulk data (zero fraction) or reference data (novel cell type). **a** Performance when pseudo-bulk samples lack a given cell type. "Zero_immune" pseudo-bulk mixtures were generated as described in the partial reference simulation but with no immune cells; "Regular" mixtures match those in the partial reference simulation. Left: 40 samples with matched references; right: 378 samples without matched references. Bars show Lin's concordance correlation coefficients (CCC) between true and estimated fractions for the five present cell types, with mean estimated immune cell fractions annotated above. **b** Deconvolution with no immune-cell reference. All 418 pseudo-bulk samples were deconvolved using 40 randomly selected references for the other five cell types only. Scatter plots compare true fractions (x-axis) versus estimated fractions (y-axis). Each panel reports CCC on the top-left corner. astro, astrocyte; inh, inhibitory neuron; OPC, oligodendrocyte progenitor cell; oligo, oligodendrocyte; ex, excitatory neuron

used Pearson's correlation and not the CCC because RNAScope/IF fractions summed up to 0.74 on average (which is much less than 1), suggesting we can only estimate the correlation between estimates and RNAScope/IF proportions up to potential location and scale shifts. Correlations between each method's estimated and RNAScope/IF fractions for each RNA extraction and library preparation combination are given in Fig. 5. BLEND accurately estimates cellular fractions across all data preparation conditions, outperforming all competing methods in five of the six conditions.

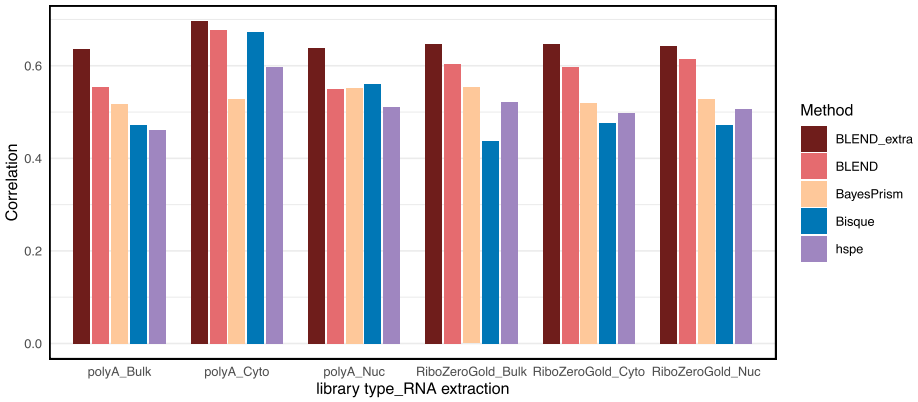

Fig. 5 Benchmarking cellular deconvolution methods with real human DLPFC data. Pearson's correlation coefficients are calculated across cell types and bulk samples under different library types and RNA extraction combinations [19]. MuSiC is not presented because of negative correlations in most settings. BLEND_extra is the setting where BLEND additionally used the Sutton reference list

We next sought to determine whether we could improve BLEND's performance by simply providing it with more references. Sutton et al. [11] collected nine brain datasets and provided a list of corresponding CTS reference matrices [23–31] with 6388 genes. These datasets were from varied sources and sequenced by different technologies. Reference names and original data sources are in Additional file 1: Table S1. Without any data evaluation and processing, we added these references to the reference list and re-ran BLEND. These extra references significantly improved BLEND's performance across all conditions ("BLEND_extra" in Fig. 5). These results suggest BLEND can leverage a large set of references to better estimate each subject's CTS expression to improve deconvolution results.

BLEND detects cell fraction changes in Alzheimer's disease

To demonstrate BLEND's ability to enhance downstream analyses, we used BLEND to detect cellular abundance changes in Alzheimer's disease (AD). The Mount Sinai Brain Bank (MSBB) study [32] provided bulk RNA-seq data of 894 samples collected from postmortem human brain, 850 of which had Braak AD-staging score (bbscore) for progression of neurofibrillary neuropathology. With tissues collected from four brain regions, Brodmann area information was also provided. To illustrate the adaptability of BLEND's automated reference selection, we used the reference list provided by Sutton et al. [11] that we also used in the real DLPFC data benchmarking above.

We assessed the relationship between bbscore and cellular fractions for three major AD-related cell types: excitatory neurons, inhibitory neurons, and microglia. As shown in Fig. 6a–c, BLEND's estimates suggest excitatory and inhibitory neuron fractions decrease and microglia fractions increase with increasing bbscore, which aligns with existing literature [33–35].

We next considered whether the above associations could also be identified if we used a different method to estimate cell fractions. We were specifically interested in variation due to the user-specified reference, which is a critical factor in many existing methods. We therefore repeated the analysis after using BayesPrism or hspe with one of the

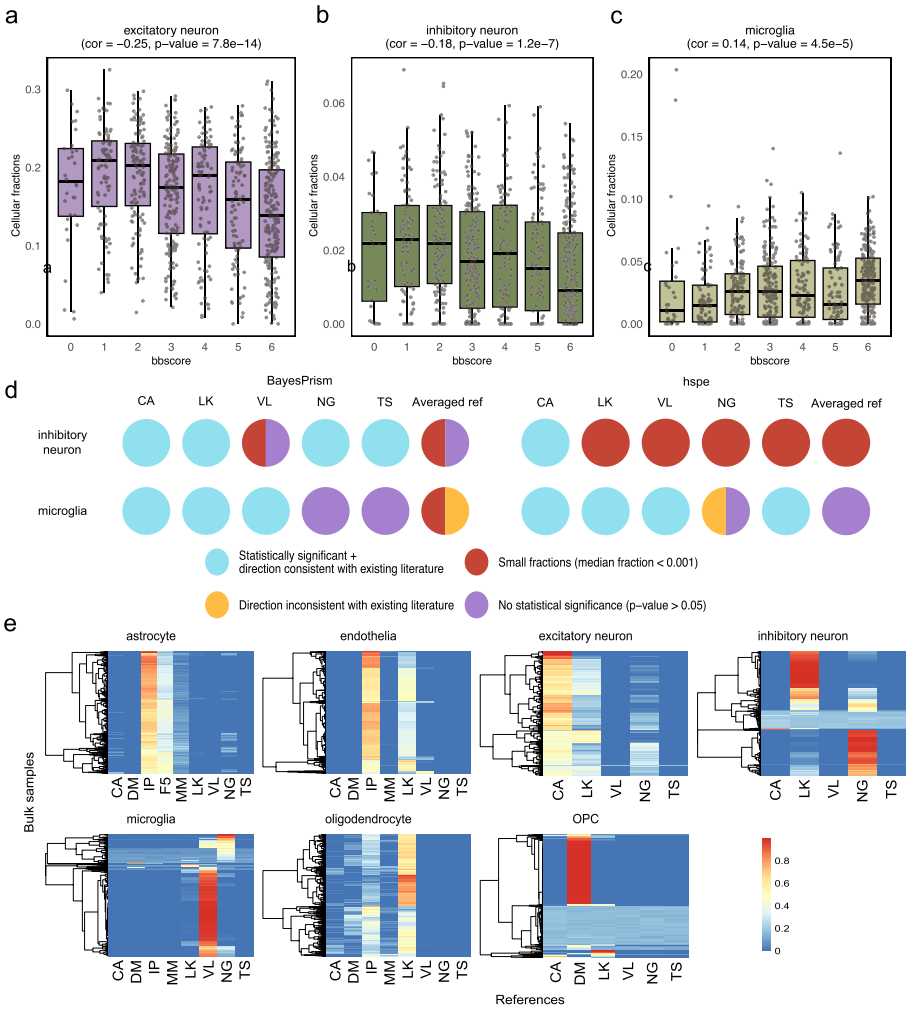

Fig. 6 Analyzing MSBB postmortem human brain data. **a**–**c** Differential cellular fractions with AD progression in three cell types. The x-axes represent Braak AD-staging score (bbscore), and the y-axes represent cellular fractions. Pearson's correlation test results between bbscore and cellular fractions are shown in brackets below titles. **d** Performance of BayesPrism and hspe when being used to repeat differential analyses using different single references and the averaged reference. The differential analyses of excitatory neurons are reasonable in all method-reference settings. The colors of the pie charts represent issues the corresponding settings have. **e** Heatmap of reference mixing proportions. Rows represent bulk samples, and columns represent references

references used by BLEND or with the averaged reference to estimate cell fractions. We only consider BayesPrism and hspe here because these methods allow for single reference matrix input (a CTS expression profile matrix). As shown in Fig. 6d, results for inhibitory and microglia depend heavily on the choice of reference, where most references suffer from issues including (1) small fraction estimates (median fraction < 0.001, given all considered cell types are common), (2) an AD association whose direction is inconsistent with the current literature, and (3) no statistical significance (p value > 0.05). BLEND avoids such issues related to manual reference selection by using the data to learn the most appropriate reference for each subject.

BLEND's estimated reference mixing proportions can help us understand the importance of reference selection. While reference mixing proportions of bulk samples from

unrelated subjects were estimated independently, heatmaps in Fig. 6e all indicate overtly non-random patterns. Several factors appear to dictate reference selection. First, reference quality appears to matter the most. Astrocytes and endothelial cells both preferred the IP reference, which was created using acutely purified human astrocytes and other major cell types [29], thereby serving as a high-quality reference. Excitatory neurons selected the CA reference, which is consistent with Hodge et al. [23] only being able to validate the excitatory neuron label from this reference. OPCs preferred the DM reference, which was derived from surgically resected brain tissues [24] that generally have higher quality than frozen tissues. Second, when two references are of similar quality, biological factors appear to become relevant. We found a significant association between the Brodmann area of bulk samples and the reference preference for inhibitory neurons (chi-squared test p value $< 2.2 \times 10^{-16}$, Methods). The NG reference measured human brain samples from Brodmann area 9 [26], and 45.4% of bulk samples choosing NG were from the adjacent Brodmann area 10. The LK reference includes information from Brodmann areas responsible for human executive functions [28], where 44.0% of bulk samples choosing LK were sampled from area 44, which is located in this functional region.

Discussion

In summary, we presented BLEND, a novel hierarchical Bayesian model for cellular deconvolution that accurately and robustly estimates cellular fractions by using the data to learn each subject's most appropriate reference. As far as we are aware, our method is the first to automate reference selection systematically. By customizing references for bulk samples through exploring convex hulls of all the available references, BLEND accommodates cross-sample, cross-data, and cross-technology heterogeneity to improve cellular fraction estimation. We used realistic pseudo-bulk data and real multi-assay data from postmortem human brains to comprehensively validate BLEND and show that it significantly outperforms existing methods. By deconvolving large-scale human brain data, we detected cellular abundance changes with AD progression, showcasing its potential in data integration and real-life application.

BLEND also exhibits robust performance despite cell type missingness in bulk data or reference data, where we proposed a surrogate uniform reference approach for the novel cell type. In practice, we recommend including the surrogate reference when the presence of an unknown cell type is uncertain. If the estimated mean fraction of that novel cell type falls below 0.01, it can be safely omitted; otherwise, it should be retained.

Notably, BLEND demonstrated its capability to find the most suitable references for bulk samples automatically, eliminating the need for reference quality evaluation. Moreover, BLEND needs no data transformation/normalization or cell type marker gene selection before deconvolution. All of these facts illustrate that BLEND is an accurate, robust, and easy-to-use deconvolution tool.

Nevertheless, there are some drawbacks of BLEND. First, the computational time of our Gibbs sampler and EM-MAP algorithms can be relatively long, especially when all genes are used and hundreds of references are provided to BLEND. However, the EM-MAP algorithm has been implemented using Rcpp and allows parallelization, which accelerates computation enormously. Second, like existing deconvolution methods, BLEND assumes genes are independent conditional on cellular fractions, which is likely

incorrect. While the veracity of this assumption does not impact point estimates, it does suggest credible and confidence intervals for true fractions will be too narrow. It should be noted, however, that while we cannot quantify the error in BLEND's estimates, this error does not compromise downstream applications for two reasons. First, the error is mean-zero and independent of biological variables (e.g., bbscore in the Alzheimer's disease application), meaning correlations between fraction estimates and biological variables will remain accurate. Second, the error is small because BLEND estimates fractions using thousands of genes.

BLEND serves as an accurate and robust tool for cell type-specific analyses. In the future, it will be interesting to design a unified statistical tool that can perform gene-gene correlation-aware deconvolution utilizing multiple references and incorporate cellular fraction estimate uncertainties in differential analysis.

Conclusions

In summary, we present BLEND, a probabilistic cellular deconvolution model with individualized cell type reference integration. By explicitly modeling the most suitable references for each bulk RNA-seq sample, BLEND accommodates cross-sample, cross-data, and cross-technology heterogeneity to substantially improve deconvolution accuracy compared to state-of-the-art methods. Furthermore, BLEND does not require data normalization/transformation or cell type marker gene selection, and thus serves as a user-friendly computational tool. In our Alzheimer's disease application, we showed BLEND avoided issues related to manual reference selection and was able to identify cellular changes with disease progression. These features, together with our easy-to-use software, make BLEND a robust and accurate deconvolution tool for use in transcriptomic analysis pipelines.

Methods

Notation used throughout the manuscript

For any positive integer n, we let $[n] = \{1, \ldots, n\}$. For $\boldsymbol{p} \in [0,1]^n$ satisfying $\sum_{i=1}^n p_i = 1$, we call the random variable $x \in [n]$ categorically distributed as $x \sim \mathrm{Cat}([n], \boldsymbol{p})$ if $\mathbb{P}(x = i) = p_i$ for all $i \in [n]$.

Overview of BLEND's probability model

BLEND models the deconvolution of bulk count RNA-seq data using multiple CTS gene expression references. Let $X_{n,g} \in \mathbb{N}_0$ be the bulk RNA-seq count for gene $g \in [G]$ in subject $n \in [N]$ and $R_n = \sum_{g=1}^G X_{n,g}$ be the library size for subject n. Let $\boldsymbol{\mu}_n \in [0,1]^T$ be the nth subject's vector of cell type fractions across T cell types, which satisfies $\sum_{t=1}^T \mu_{n,t} = 1$. We assume there are M_t available references for cell type t and let $\boldsymbol{\Phi}_{t,m} \in [0,1]^G$ be the mth reference for cell type t. As this is a distribution over genes, $\sum_{g=1}^G \Phi_{t,m,g} = 1$. Our goal is to estimate cell type proportions $\boldsymbol{\mu}_n$ given the observed data $\{X_{n,g}\}_{g\in[G]}$ and references $\{\boldsymbol{\Phi}_{t,m}\}_{t\in[T];\, m\in[M_t]}$.

BLEND uses a generative "bag-of-words" model [12] to generate each read's gene identity. Briefly, we first draw the read's cell type of origin. The probability that the read is assigned to gene g is then determined by the originating cell type's reference. Unlike traditional LDA, which assumes references are the same across subjects

[12], BLEND allows references to be subject-specific by assuming that the reference for cell type t in subject n lies in the convex hull of $\{\boldsymbol{\Phi}_{t,m}\}_{m\in[M_t]}$. We formalize these ideas in the following generative probability model, which generates the observed data $\{X_{n,g}\}_{g\in[G]}$ for each subject n. The graphical plate representation is in Additional file 1: Fig. S1. We implicitly condition on library size R_n and observed references $\boldsymbol{\Phi}_{t,m}$, although we leave them out of conditioning arguments for presentation simplicity.

(i) Generate a length-T vector of cell type proportions $\boldsymbol{\mu}_n \mid \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha})$.
(ii) For each cell type $t \in [T]$, generate a length-M_t vector of reference mixing proportions $\boldsymbol{\psi}_{n,t} \mid \boldsymbol{\beta}_t \sim \text{Dirichlet}(\boldsymbol{\beta}_t)$.
(iii) For read $r \in [R_n]$, draw its cell type source $Y_{n,r} \mid \boldsymbol{\mu}_n \sim \text{Cat}([T], \boldsymbol{\mu}_n)$.
(iv) Draw the rth read's gene identity, $\tilde{X}_{n,r}$, as $\tilde{X}_{n,r} \mid (Y_{n,r} = t, \boldsymbol{\psi}_{n,t}) \sim \text{Cat}([G], \boldsymbol{\Phi}_{n,t}^{\text{BLEND}})$, where $\boldsymbol{\Phi}_{n,t}^{\text{BLEND}} = \sum_{m=1}^{M_t} \psi_{n,t,m}\boldsymbol{\Phi}_{t,m}$.
(v) For each gene $g \in [G]$, set $X_{n,g} = \sum_{r=1}^{R_n} 1\{\tilde{X}_{n,r} = g\}$.

Step (iv) lets each subject have their own reference $\boldsymbol{\Phi}_{n,t}^{\text{BLEND}}$ for each cell type t, which is assumed to be a convex combination of the observed references $\{\boldsymbol{\Phi}_{t,m}\}_{m\in[M_t]}$. This is quite general, as it allows a subject's reference to be the average of observed references ($\psi_{n,t,m} = 1/M_t$ for all m) or be determined by only a subset of references ($\psi_{n,t,m}$ is large for some references m and small for others). The latter will occur when references are sampled from multiple populations, where $\psi_{n,t,m}$ will be large if reference m and the CTS expression for subject n are sampled from the same population. This ensures our model captures the variation in CTS expression across subjects.

Parameter estimation strategies of BLEND

Here we derive a Gibbs sampler to sample from the posterior $\mathbb{P}(\boldsymbol{\mu}_n, \{\boldsymbol{\psi}_{n,t}\}_{t\in[T]} \mid \{X_{n,g}\}_{g\in[G]}, \boldsymbol{\alpha}, \{\boldsymbol{\beta}_t\}_{t\in[T]})$, where our estimator for cell type proportions is then the posterior expectation of $\boldsymbol{\mu}_n$. To ensure Gibbs updates are tractable, we introduce a new latent variable by noting that drawing the rth read's gene identity in step (iv) of the above generative probability model is equivalent to the following two-step procedure:

$$Z_{n,r}|(Y_{n,r} = t, \boldsymbol{\psi}_{n,t}) \sim \text{Cat}([M_t], \boldsymbol{\psi}_{n,t})$$
$$\tilde{X}_{n,r}|(Y_{n,r} = t, Z_{n,r} = m) \sim \text{Cat}([G], \boldsymbol{\Phi}_{t,m}).$$

The new latent variable $Z_{n,r}$ is interpretable as the rth read's reference source. To derive the Gibbs sampler, let

$$S_{n,g,t,m} = \sum_{r=1}^{R_n} 1\{\tilde{X}_{n,r} = g, Y_{n,r} = t, Z_{n,r} = m\}$$

be the number of gene g's reads that are drawn from cell type t using reference $m \in [M_t]$, and define $\boldsymbol{S}_{n,g} = \{S_{n,g,t,m}\}_{t\in[T];m\in[M_t]}$. In Gibbs sampling, we sample from marginal posteriors:

$$\mathbb{P}(\{\boldsymbol{S}_{n,g}\}_{g\in[G]} \mid \boldsymbol{\mu}_n, \{\boldsymbol{\psi}_{n,t}\}_{t\in[T]}, \{X_{n,g}\}_{g\in[G]}, \boldsymbol{\alpha}, \{\boldsymbol{\beta}_t\}_{t\in[T]}),$$
$$\mathbb{P}(\boldsymbol{\mu}_n \mid \{\boldsymbol{S}_{n,g}\}_{g\in[G]}, \{\boldsymbol{\psi}_{n,t}\}_{t\in[T]}, \{X_{n,g}\}_{g\in[G]}, \boldsymbol{\alpha}, \{\boldsymbol{\beta}_t\}_{t\in[T]}),$$
$$\mathbb{P}(\{\boldsymbol{\psi}_{n,t}\}_{t\in[T]} \mid \{\boldsymbol{S}_{n,g}\}_{g\in[G]}, \boldsymbol{\mu}_n, \{X_{n,g}\}_{g\in[G]}, \boldsymbol{\alpha}, \{\boldsymbol{\beta}_t\}_{t\in[T]}).$$

The Gibbs sampler converges quickly due to the large library sizes in bulk RNA-seq data. We let the entries of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_t$ be 0.01 in practice, which ensures the priors for $\boldsymbol{\mu}_n$ and $\boldsymbol{\psi}_{n,t}$ are non-informative.

While our Gibbs sampler is relatively fast and easy to implement, we would ideally optimize computation to be able to estimate cell-type proportions in modern datasets that include thousands of bulk samples. We therefore present an algorithm to maximize the posterior distribution, that is, to derive the maximum a posteriori (MAP) estimate for $\boldsymbol{\mu}_n$ and $\{\boldsymbol{\psi}_{n,t}\}_{t\in[T]}$. We use an expectation-maximization (EM) algorithm to maximize the log-posterior for $\boldsymbol{\mu}_n$ and $\{\boldsymbol{\psi}_{n,t}\}_{t\in[T]}$ conditional on the data $\{X_{n,g}, \boldsymbol{S}_{n,g}\}_{g\in[G]}$ and Dirichlet hyperparameters $\boldsymbol{\alpha}, \{\boldsymbol{\beta}_t\}_{t\in[T]}$. Briefly, we consider the latent variables $\boldsymbol{S}_{n,g} = \{S_{n,g,t,m}\}_{t\in[T];m\in[M_t]}$ and define the latent variable-augmented log-posterior to be

$$\tilde{l}(\boldsymbol{\mu}_n, \{\boldsymbol{\psi}_{n,t}\}_{t\in[T]}) = \log \mathbb{P}(\boldsymbol{\mu}_n, \{\boldsymbol{\psi}_{n,t}\}_{t\in[T]} \mid \{X_{n,g}, \boldsymbol{S}_{n,g}\}_{g\in[G]}, \boldsymbol{\alpha}, \{\boldsymbol{\beta}_t\}_{t\in[T]}).$$

The E-step of the EM algorithm is then:

$$Q(\boldsymbol{\mu}_n, \{\boldsymbol{\psi}_{n,t}\}_{t\in[T]}) = E\{\tilde{l}(\boldsymbol{\mu}_n, \{\boldsymbol{\psi}_{n,t}\}_{t\in[T]})\},$$

where expectation is taken with respect to the distribution $\mathbb{P}(\{\boldsymbol{S}_{n,g}\}_{g\in[G]} \mid \{X_{n,g}\}_{g\in[G]}, \boldsymbol{\mu}_n^{(0)}, \{\boldsymbol{\psi}_{n,t}^{(0)}\}_{t\in[T]})$, where $\boldsymbol{\mu}_n^{(0)}$ and $\boldsymbol{\psi}_{n,t}^{(0)}$ are the algorithm's current values of $\boldsymbol{\mu}_n$ and $\boldsymbol{\psi}_{n,t}$. The M-step is

$$\text{argmax}_{\boldsymbol{\mu}_n, \{\boldsymbol{\psi}_{n,t}\}_{t\in[T]}} \; Q(\boldsymbol{\mu}_n, \{\boldsymbol{\psi}_{n,t}\}_{t\in[T]}),$$

subject to $\boldsymbol{\mu}_n \geq \mathbf{0}$, $\boldsymbol{\psi}_{n,t} \geq \mathbf{0}$, $\sum_{t=1}^{T} \mu_{n,t} = 1$, and $\sum_{m=1}^{M_t} \psi_{n,t,m} = 1$ for all $t \in [T]$.

Aside from being able to provide the same parameter estimates, EM-MAP simplifies estimation by replacing the time-consuming sampling process and the need for additional samples after burn-in with purely algebraic operations and direct parameter estimation upon convergence. We implement the Gibbs sampler in R and the EM-MAP algorithm in Rcpp. In the MSBB application, it takes the Gibbs sampler 36 min to deconvolve one bulk sample (6388 genes) using nine references. Apart from providing consistent estimates (Additional file 1: Fig. S2), the EM-MAP algorithm significantly reduces computation time to 1.5 min.

Cell size adjustment

After the parameter estimation, we focus on the interpretation and use of the cellular fraction parameter $\boldsymbol{\mu}_n$. In BLEND, we model RNA transcripts directly. Thus, $\boldsymbol{\mu}_n$ should be interpreted as proportions of transcript attributed to cell types in the tissue (RNA fractions). When cell fractions are of interest, we need to consider the varying abundance of transcripts of different cell types, which is quantified by a cell size vector $\boldsymbol{S} \in \mathbb{R}_+^T$. $\boldsymbol{S}$ can be given or estimated by the average library sizes of cell types, the latter of which was used in our experiments. Then, we estimate cell fractions by

$$\mu_{n,t}^{cell} = \frac{\mu_{n,t}/S_t}{\sum_{t*=1}^{T}(\mu_{n,t*}/S_{t*})}.$$

In benchmarking studies with reliable ground truth of cell fractions, such as simulation, we perform cell size adjustment first before comparison.

Chi-squared test in the MSBB data analysis

We tested the independence between Brodmann areas and the reference preferences of bulk samples using the chi-squared test on the MSBB data. For cell type t, we estimated the reference mixing proportions $\hat{\boldsymbol{\psi}}_t = (\hat{\boldsymbol{\psi}}_{1,t}, \hat{\boldsymbol{\psi}}_{2,t}, \ldots, \hat{\boldsymbol{\psi}}_{N,t})'$. We then defined the preferred reference for a bulk sample-cell type pair (n, t) as the reference with the highest value in $\hat{\boldsymbol{\psi}}_{n,t}$. We constructed a contingency table using reference preferences and Brodmann areas of bulk samples and performed a chi-squared test. The p value $< 2.2 \times 10^{-16}$ indicated major statistical significance that was below the computation limit in R.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03765-6.

Additional file 1. Supplementary Notes: Technical details of BLEND, Tables S1-S6, Fig. S1-S6 referenced in the main text.

Acknowledgements
The results published here are in whole or in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org). 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, R01AG36042 (5hC methylation, ATACseq), RC2AG036547 (H3K9Ac), R01AG36836 (RNAseq), R01AG48015 (monocyte RNAseq), RF1AG57473 (single nucleus RNAseq), U01AG32984 (genomic and whole exome sequencing), U01AG46152 (ROSMAP AMP-AD, targeted proteomics), U01AG46161 (TMT proteomics), U01AG61356 (whole genome sequencing, targeted proteomics, ROSMAP AMP-AD), the Illinois Department of Public Health (ROSMAP), and the Translational Genomics Research Institute (genomic). Additional phenotypic data can be requested at www.radc.rush.edu. The MSBB data were generated from postmortem brain tissue collected through the Mount Sinai VA Medical Center Brain Bank and were provided by Dr. Eric Schadt from Mount Sinai School of Medicine.

Peer review information
Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
Under the supervision of J.W. and C.M., P.H. conceived the idea, derived algorithms, and conducted experiments. P.H., J.W., and C.M. designed the experiments and wrote the manuscript. P.H. and M.C. implemented the software. J.W. provided funding support. All authors read and approved the final manuscript.

Funding
This research was funded in part through NIH's R01AG080590 and R03OD034501.

Data availability
The datasets supporting the conclusions of this article are publicly available: Mathys data (https://www.synapse.org/#!Synapse:syn52293417) [36], Fujita data (https://www.synapse.org/#!Synapse:syn31512863) [37], Huuki-Myers real human DLPFC benchmarking data (https://github.com/LieberInstitute/Human_DLPFC_Deconvolution) [38], MSBB bulk RNA-seq data (https://www.synapse.org/Synapse:syn22025006) [39], and Sutton reference list (https://github.com/Voineagulab/BrainCellularComposition) [40]. The accompanied R package is available at https://github.com/Penghuihuang2000/BLEND [41] under GPL (>= 3) license. Source code related to experiments is available at Zenodo https://zenodo.org/records/15237115 [42].

Declarations

Ethics approval and consent to participate
Not applicable for this study.

Competing interests
The authors declare no competing interests.

Received: 22 October 2024 Accepted: 31 August 2025
Published online: 02 October 2025

References

1. Thind AS, Monga I, Thakur PK, Kumari P, Dindhoria K, Krzak M, et al. Demystifying emerging bulk RNA-seq applications: the application and utility of bioinformatic methodology. Brief Bioinform. 2021;22(6):bbab259.
2. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. 2014;15(2):R31.
3. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, et al. Cell type-specific gene expression differences in complex tissues. Nat Methods. 2010;7(4):287–9.
4. Wang J, Devlin B, Roeder K. Using multiple measurements of tissue to estimate subject-and cell-type-specific gene expression. Bioinformatics. 2020;36(3):782–8.
5. Wang J, Roeder K, Devlin B. Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. Genome Res. 2021;31(10):1807–18.
6. Hu M, Chikina M. Heterogeneous pseudobulk simulation enables realistic benchmarking of cell-type deconvolution methods. Genome Biol. 2024;25(1):169.
7. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat Commun. 2020;11(1):5650.
8. Schelker M, Feau S, Du J, Ranu N, Klipp E, MacBeath G, et al. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nat Commun. 2017;8(1):2032.
9. Cai M, Yue M, Chen T, Liu J, Forno E, Lu X, et al. Robust and accurate estimation of cellular fraction from tissue omics data via ensemble deconvolution. Bioinformatics. 2022;38(11):3004–10.
10. Huang P, Cai M, Lu X, McKennan C, Wang J. Accurate estimation of rare cell-type fractions from tissue omics data via hierarchical deconvolution. Ann Appl Stat. 2024;18(2):1178–94.
11. Sutton GJ, Poppe D, Simmons RK, Walsh K, Nawaz U, Lister R, et al. Comprehensive evaluation of deconvolution methods for human brain gene expression. Nat Commun. 2022;13(1):1358.
12. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3(Jan):993–1022.
13. Mathys H, Peng Z, Boix CA, Victor MB, Leary N, Babu S, et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. Cell. 2023;186(20):4365–85.
14. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious orders study and rush memory and aging project. J Alzheimers Dis. 2018;64(S1):S161-89.
15. Chu T, Wang Z, Pe'er D, Danko CG. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. Nat Cancer. 2022;3(4):505–17.
16. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019;10(1):380.
17. Jew B, Alvarez M, Rahmani E, Miao Z, Ko A, Garske KM, et al. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nat Commun. 2020;11(1):1971.
18. Hunt GJ, Gagnon-Bartsch JA. The role of scale in the estimation of cell-type proportions. Ann Appl Stat. 2021;15(1):270–86.
19. Huuki-Myers LA, Montgomery KD, Kwon SH, Cinquemani S, Eagles NJ, Gonzalez-Padilla D, et al. Benchmark of cellular deconvolution methods using a multi-assay dataset from postmortem human prefrontal cortex. Genome Biol. 2025;26(1):1-35.
20. Lin LIK. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45(1):255–68.
21. Fujita M, Gao Z, Zeng L, McCabe C, White CC, Ng B, et al. Cell subtype-specific effects of genetic variation in the Alzheimer's disease brain. Nat Genet. 2024;56(4):605–14.
22. Caggiano C, Celona B, Garton F, Mefford J, Black BL, Henderson R, et al. Comprehensive cell type decomposition of circulating cell-free DNA with CelFiE. Nat Commun. 2021;12(1):2717.
23. Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, et al. Conserved cell types with divergent features in human versus mouse cortex. Nature. 2019;573(7772):61–8.
24. Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, et al. A survey of human brain transcriptome diversity at the single cell level. Proc Natl Acad Sci U S A. 2015;112(23):7285–90.
25. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. Nature. 2014;507(7493):462–70.
26. Nagy C, Maitra M, Tanti A, Suderman M, Théroux JF, Davoli MA, et al. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. Nat Neurosci. 2020;23(6):771–81.
27. Velmeshev D, Schirmer L, Jung D, Haeussler M, Perez Y, Mayer S, et al. Single-cell genomics identifies cell type-specific molecular changes in autism. Science. 2019;364(6441):685–9.
28. Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol. 2018;36(1):70–80.
29. Zhang Y, Sloan SA, Clarke LE, Caneda C, Plaza CA, Blumenthal PD, et al. Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. Neuron. 2016;89(1):37–53.
30. Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S, et al. An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. J Neurosci. 2014;34(36):11929–47.
31. Tasic B, Yao Z, Graybuck LT, Smith KA, Nguyen TN, Bertagnolli D, et al. Shared and distinct transcriptomic cell types across neocortical areas. Nature. 2018;563(7729):72–8.

32. Wang M, Beckmann ND, Roussos P, Wang E, Zhou X, Wang Q, et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Sci Data. 2018;5(1):1–16.
33. Fu H, Rodriguez GA, Herman M, Emrani S, Nahmani E, Barrett G, et al. Tau pathology induces excitatory neuron loss, grid cell dysfunction, and spatial memory deficits reminiscent of early Alzheimer's disease. Neuron. 2017;93(3):533–41.
34. Verret L, Mann EO, Hang GB, Barth AM, Cobos I, Ho K, et al. Inhibitory interneuron deficit links altered network activity and cognitive dysfunction in Alzheimer model. Cell. 2012;149(3):708–21.
35. Hansen DV, Hanson JE, Sheng M. Microglia in Alzheimer's disease. J Cell Biol. 2018;217(2):459–72.
36. Mathys H, Peng Z, Boix CA, Victor MB, Leary N, Babu S, et al. The MIT ROSMAP Single-Nucleus Multiomics Study (MITROSMAPMultiomics) [Datasets]. Synapse. https://www.synapse.org/Synapse:syn52293417. Accessed 2023.
37. Fujita M, Gao Z, Zeng L, McCabe C, White CC, Ng B, et al. Gene expression (snRNAseq – DLPFC, experiment 2) [Datasets]. Synapse. https://www.synapse.org/Synapse:syn31512863. Accessed 2023.
38. Huuki-Myers LA, Montgomery KD, Kwon SH, Cinquemani S, Eagles NJ, Gonzalez-Padilla D, et al. Human DLPFC deconvolution [Datasets]. GitHub. https://github.com/LieberInstitute/Human_DLPFC_Deconvolution. Accessed 2024.
39. Wang M, Beckmann ND, Roussos P, Wang E, Zhou X, Wang Q, et al. Mount Sinai Brain Bank (MSBB) study [Datasets]. Synapse. https://www.synapse.org/Synapse:syn22025006. Accessed 2022.
40. Sutton GJ, Poppe D, Simmons RK, Walsh K, Nawaz U, Lister R, et al. BrainCellularComposition [Datasets]. GitHub. https://github.com/Voineagulab/BrainCellularComposition. Accessed 2024.
41. Huang P, Cai M, McKennan C, Wang J. BLEND R package. GitHub. 2025. https://github.com/Penghuihuang2000/BLEND.
42. Huang P, Cai M, McKennan C, Wang J. BLEND source code. Zenodo. 2025. https://doi.org/10.5281/zenodo.15237115.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.