BLADE: Bayesian Log-normAl DEconvolution for enhanced in silico microdissection of bulk gene expression data

Bárbara Andrade Barbosa
Amsterdam UMC
Saskia van Asten
Amsterdam UMC https://orcid.org/0000-0001-6498-1176
Ji-won Oh
Kyungpook National University
Arantza Fariña-Sarasqueta
Amsterdam University Medical Center
Joanne Verheij
Amsterdam University Medical Center
Frederike Dijk
Amsterdam UMC
Hanneke van Laarhoven
Department of Medical Oncology, Cancer Center Amsterdam, Amsterdam University Medical Center
Bauke Ylstra
Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Pathology, Cancer Center Amsterdam, Amsterdam https://orcid.org/0000-0001-9479-3010
Juan Garcia-Vallejo
Amsterdam UMC
Mark van de Wiel
Amsterdam UMC
Yongsoo Kim (yo.kim@amsterdamumc.nl)
Amsterdam UMC, location VUmc https://orcid.org/0000-0002-2995-2131

Article

Keywords: bulk gene expression, Bayesian Log-normAl Deconvolution

Posted Date: December 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-123595/v1

Version of Record: A version of this preprint was published at Nature Communications on October 20th, 2021. See the published version at https://doi.org/10.1038/s41467-021-26328-2.

BLADE: Bayesian Log-normAl DEconvolution for enhanced *in silico* microdissection of bulk gene expression data

Bárbara Andrade Barbosa[1], Saskia D. van Asten[1,2], Ji-won Oh[3,4], Arantza Farina Sarasqueta[5], Joanne Verheij[5], Frederike Dijk[5], Hanneke van Laarhoven[6], Bauke Ylstra[1], Juan Garcia Vallejo[2], Mark van de Wiel[7*], Yongsoo Kim[1*]

[1]Department of Pathology, Cancer-Center Amsterdam, Amsterdam UMC location VUmc, 1081 HV, Amsterdam, the Netherlands

[2]Department of Molecular Cell Biology & Immunology, Amsterdam UMC location VUmc, Amsterdam Infection and immunity Institute, 1081 HZ, Amsterdam, the Netherlands

[3]Department of Anatomy, School of Medicine, Kyungpook National University, 41940, Daegu, South Korea

[4]Bio-Medical Research Institute, Kyungpook National University Hospital, 41944, Daegu, South Korea

[5]Department of Pathology, Amsterdam UMC location AMC, 1105 AZ, Amsterdam, the Netherlands

[6]Department of Medical Oncology, Cancer Center Amsterdam, Amsterdam UMC location AMC, 1105 AZ, Amsterdam, the Netherlands

[7]Department of Epidemiology and Biostatistics, Amsterdam UMC location VUmc, 1081 HV, Amsterdam, the Netherlands

* To whom correspondence should be addressed. Email: yo.kim@amsterdamumc.nl Email: mark.vdwiel@amsterdamumc.nl

Abstract (145/150)

High-resolution deconvolution of bulk gene expression profiles is pivotal to characterize the complex cellular make-up of tissues, such as tumor microenvironment. Single-cell RNA-seq provides reliable prior knowledge for deconvolution, however, a comprehensive statistical model is required for efficient utilization due to the inherently variable nature of gene expression. We introduce BLADE (Bayesian Log-normAl Deconvolution), a comprehensive probabilistic framework to estimate both cellular make-up and gene expression profiles of each cell type in each sample. Unlike previous comprehensive statistical approaches, BLADE can handle >20 cell types thanks to the efficient variational inference. Throughout an intensive evaluation using >700 datasets, BLADE showed enhanced robustness against gene expression variability and better completeness than conventional methods, in particular to reconstruct gene expression profiles of each cell type. All-in-all, BLADE is a powerful tool to unravel heterogeneous cellular activity in complex biological systems based on standard bulk gene expression data.

Introduction

Over the past decade, gene expression profiling has been applied to elucidate the complexity of transcriptional regulation in diverse biological contexts, such as cancer[1,2]. Conventional gene expression profiling, based on either RNA sequencing (RNA-seq) or microarrays, captures cumulative gene expression levels of many cells combined. It is therefore often referred to as bulk gene expression profiling, in order to distinguish it from the recent single-cell gene expression profiling technologies[3]. In oncology, single-cell RNA sequencing (scRNA-seq) is employed to study cellular heterogeneity within a tumor, composed of malignant (tumor) and non-malignant cells[4–10]. However, scRNA-seq has serious limitations: apart from technical challenges such as drop-out, only a limited number

of samples can be profiled due to the high cost and technical difficulties[11,12] which altogether hinder its application to large series and translation to clinical applications.

Several computational deconvolution methods have been developed to predict cellular composition from bulk RNA-seq by employing a signature of pre-determined cell type-specific gene expression profiles. Initially, these signatures were constructed by sorting each cell type followed by gene expression profiling[13], whereas recent methods such as CIBERSORTx[14] and MuSiC[15] employed scRNA-seq data for this purpose. The majority of approaches perform linear regression to reconstruct the bulk gene expression profiles using the gene expression signatures, where the regression coefficients correspond to the cellular composition. However, the standard regression approach does not account for variability in gene expression within the same cell type, and may therefore render biased results.

To the best of our knowledge, there is no deconvolution method that can adequately and also efficiently account for the variability of gene expression within the same cell type. Modelling gene expression variability is challenging specifically for deconvolution due to the incompatibility of the log-normalization[16] which significantly stabilizes gene expression variability. Without the log-normalization (i.e., in linear-scale), gene expression data has a heavily skewed distribution, which is not properly accounted for by the standard linear regression approaches. Currently, there are few probabilistic deconvolution approaches that take skewed variability into account, but these methods handle only a restricted number of cell types due to difficulties in optimization (e.g. three cell types in DeClust[17] and Demix/DemixT[18]). Recently, CIBERSORTx introduced a two-step approach to address variable gene expression profiles across the samples: first estimate cellular fraction (deconvolution) and then reconstruct gene expression per cell type in each sample (purification). However, for some genes, the purification step of CIBERSORTx is an underdetermined optimization problem with infinite number of solutions, which leads to incomplete purification that excludes those genes.

Here, we introduce BLADE (Bayesian Log-normAl DEconvolution), a novel Bayesian method that jointly performs deconvolution and purification in a single-step, taking into account prior knowledge of cell type specific gene expression profiles obtained from scRNA-seq data. BLADE takes a Bayesian framework that integrates two signatures of mean and variability of gene expression per-cell type using a log-normal probability model. The unified probabilistic model for both deconvolution and purification of BLADE can leverage the prior knowledge for purification as well, which may remedy the underdetermination issue. Furthermore, an efficient variational inference algorithm was developed for which we show that it can handle at least 20 cell types. Through a comprehensive evaluation based on more than 700 bulk gene expression data sets, we demonstrate a robust performance of BLADE regardless of gene expression variability. In particular, BLADE achieved high accuracy and completeness in gene expression purification, underpinning the power of the unified Bayesian framework for both tasks.

Results

Gene expression variability within a cell type

We first assessed gene expression variability within a cell type using publicly available PBMC CITE-seq data from 10x Genomics. Based on the integration and clustering analysis followed by phenotyping of 10,403 cells, we identified fifteen immune cell types, among which nine are in common, with distinct cell-surface markers and gene expression profiles (**Fig. 1a;** see **Methods** and **Supplementary Figs. S1-2**). The size of cell populations ranges from 38 regulatory T cells (0.36%) to 2,518 classical monocytes (24%). We then identified differentially expressed genes (DEGs) for each cell type. Subsequently, the standard deviation of gene expression levels per gene and per cell type was measured to assess gene expression variability among the same cell types. We identified high gene expression variability among the same cell populations, especially for DEGs without log-transformation (i.e., linear-scale; **Figs. 1b-c**). The variability further increased when cells from the two scRNA-seq datasets were combined, indicating the presence of more variability between

individuals (**Fig. 1d;** *P*<2.2x10e-16 from a paired t-test of within-sample and between-sample variability).

Modeling gene-expression variability by probabilistic distribution

To properly account for variation in gene expression, we examined multiple probability distributions. We evaluated normal distribution, negative binomial distribution, and log-normal distribution to fit the expression level of each gene per cell type without log-normalization. The normal distribution is the standard variability model in many deconvolution algorithms, including CIBERSORTx[14], EPIC[19], and ABIS[20], while the negative binomial distribution is frequently used for handling RNA-seq data[21]. The log-normal distribution is identical to the normal distribution but includes an exponential function, assuming gene expression data is normally distributed on a log-scale but not on a linear-scale. In order to evaluate the performance of these probability distributions on gene expression variability, we assessed the maximum likelihood in fitting gene expression profiles per cell type from the scRNA-seq data. The log-normal distribution, in general, shows the best performance in per-gene maximum likelihood, followed by the negative binomial and normal distributions (**Figs. 2a-b**). In particular, we noted a biased fit of the normal distribution towards outlier observations, in contrast to the log-normal and negative binomial distribution (examples in **Fig. 2c**).

We further evaluated the performance of the log-normal and negative binomial distributions in the context of deconvolution. To this end, we constructed a generic statistical deconvolution method that can model gene expression profiles with diverse probabilistic assumptions given known cellular fractions. The method approximates the convolution of random variables with an arbitrary distribution using a probabilistic generating function, for which both negative binomial and log-normal random variables can be accurately approximated (see **Methods**, **Supplementary Note 1** and **Supplementary Fig. S3**). Based on this method, we evaluated the performance of negative binomial and log-normal distribution in fitting the gene expression profiles per cell type using RNA-seq data from

TCGA. We obtained TCGA RNA-seq data of mesothelioma (TCGA-MESO; n=84) and sarcoma (TCGA-SARC; n=256), from which we estimated the fraction of eight cell types using EPIC[19], a non-probabilistic deconvolution method. Then, we applied the flexible deconvolution method with two different probabilistic assumptions, log-normal and negative binomial, to estimate expression profiles per cell type of 200 random genes. In terms of log-likelihood measured per gene, log-normal and negative binomial deconvolutions performed equally well for most of the genes, except for a few genes with a more favorable performance with log-normal (**Fig. 3**). The lower performance of the negative binomial distribution might be due to the difficulty in finding maximum likelihood parameters. Cumulatively, we concluded that the log-normal distribution is an attractive probabilistic distribution to model gene expression variability of each cell type.

Overview of BLADE: Bayesian Log-normal Deconvolution

We constructed a Bayesian Log-normal Deconvolution model, BLADE, by emulating bulk gene expression profiles through convolution of gene expression profiles per cell type (**Fig. 4a**). The bulk gene expression level of each gene j in sample i was modeled by $y_{ij} = \sum_t {f_i}^t {x_{ij}}^t + \epsilon_{ij}$. Here, the hidden variables ${f_i}^t$ and ${x_{ij}}^t$ denote the cell type t fraction for sample i and the purified cell type t gene j expression for sample i. These hidden variables ${f_i}^t$ and ${x_{ij}}^t$ are respectively endowed with the Dirichlet distribution and the log-normal distribution. To incorporate prior knowledge from scRNA-seq data, we take a hierarchical approach to model ${x_{ij}}^t$ by taking a conjugate prior of log-normal distribution with hyperparameters ${\mu^t}_{0j}$, ${\kappa^t}_{0j}$, ${\alpha^t}_{0j}$, and ${\beta^t}_{0j}$ (**Fig. 4b**). The hyperparameters are chosen based on the mean and standard deviation of each gene per cell type from the scRNA-seq data by inferring the hidden variables, we can jointly estimate the fraction of cell types, captured by ${f^t}_i$, and purified gene expression profiles of each cell type in each sample, captured by ${x^t}_{ij}$. For inference, we employed a collapsed variational inference that maximize efficiency by integrates out a subset of hidden variables with a conjugate prior in

advance. Furthermore, we employed the L-BFGS algorithm in conjunction with machine-code translated Python code for gradient and objective function calculations instead of native Python code. The compilation of native Python code by the Numba package[22] significantly accelerates gradient and objective functions that are executed thousands of times during the L-BFGS optimization (**Supplementary Fig. S4**)**.** See **Methods** and **Supplementary Note 2** for the further details of the framework. As a result, BLADE can handle many cell types (>20 cell types) and samples (>20 samples); unlike the previous log-normal based deconvolution that can account for a maximum of three cell types[18].

Robustness of BLADE deconvolution against gene expression variability.

We assessed the robustness of BLADE, CIBERSORTx, and non-negative least squares (NNLS) against gene expression variability by applying them to model-based simulation data. The simulation data was created to have diverse but controlled variability levels of gene expression profiles (standard deviation of 0.1-1.5) as well as different numbers of cell types (5-20 cell types), marker genes (100-1000 genes), and samples (5-100 samples; in total 700 training data sets). Note that NNLS is a regularized linear regression, a type of constrained linear regression used in many deconvolution methods, including EPIC[19], TIMER[23], ABIS[20], and also in the purification step of CIBERSORTx[14]. The variability levels of the simulation data were selected in order to recapitulate the range of the observed in the scRNA-seq data (up to standard deviation of 1.5 in log scale; **Fig. 1b-c**). In general, all three methods could accurately estimate cellular fractions in case of a high number of genes, a low number of cell types and a low variability level. In contrast, the performance decreased when a smaller number of genes are presented, and the number of cell types is increased (**Fig. 5a**). However, BLADE was more robust against gene expression variability. In particular, in the range of observed expression variability of differentially expressed genes in the PBMC scRNA-seq data (on average >0.5; **Fig. 1b**), BLADE significantly outperformed CIBERSORTx and NNLS.

We then compared the performance of BLADE and CIBERSORTx in estimating gene expression profiles per cell type. In this comparison, NNLS is not included because of redundancy, since the purification step of CIBERSORTx is based on NNLS. There are two modes of purification in CIBERSORTx, both of which were compared with BLADE: 1) estimating average profile per cell type across the samples (group-mode purification), and 2) estimating the profile per cell type for each sample (high-resolution-mode purification). For the data set with low variability levels, both BLADE and CIBERSORTx accurately reconstructed gene expression profiles per cell type (**Fig. 5b-c**). However, unlike BLADE, the performance of CIBERSORTx decreased rapidly as the RNA expression variability within a cell type increased. Furthermore, CIBERSORTx often excludes genes for purification, especially in high-resolution mode, when: 1) the number of cell types is larger than the number of samples, and 2) the variability in gene expression is high (**Fig. 5d-e**). BLADE could accurately estimate the gene expression profiles of each cell type in both group-mode and high-resolution mode, regardless of the number of cell types and samples without any filtering (**Fig. 5b-c**).

Application of BLADE to *in silico* mixture of PBMC scRNA-seq data

We further evaluated our method based on actual scRNA-seq data from PBMC samples that were mixed *in silico* in various known proportions to generate bulk gene expression data without any model assumption. We generated 20 bulk gene expression data sets by random sampling, followed by mixing 100 cells among the 10,403 cells from the two PBMC scRNA-seq data sets. In order to make the simulation data as realistic as possible, a cumulative sum of raw counts of 100 cells was obtained, followed by a standard normalization of the count data The resulting simulation data recapitulate the gene expression variability of 15 cell types (**Fig. 6a**; **Supplementary Fig. S5**). We constructed signature matrices for the mean and the standard deviation of 1,007 genes selected by merging the top 200

differentially expressed genes with FDR < 0.2 of each of the 15 cell types. We used the same mean signature matrix for the baseline methods, CIBERSORTx and NNLS, for a fair comparison.

BLADE outperformed CIBERSORTx and NNLS to predict the fractions for most of the 15 cell types (P-values of paired T-tests < 0.05; **Fig. 6b; Supplementary Fig. S6**). The three methods performed similarly for most cell types (Pearson correlation > 0.7), except for a few cell types with a contradictory outcome. Taking a Pearson correlation of 0.25 as a threshold, BLADE had a good predictive performance in more cell types (11 cell types) than the baseline methods, CIBERSORTx and NNLS (7 cell types). Among the 15 cell types, plasmablasts and classical/non-classical monocytes were the best predicted by all three methods, whereas the methods commonly failed to predict the composition of regulatory T-cells (Tregs), naive $CD8^+$ T-cells (NaiveCD8T), and plasmacytoid dendritic cells (pDC). These poorly predicted cell types were low abundant (less than 2%; **Supplementary Fig. S7**), indicating the difficulty in deconvolution of rare cell populations. However, some of the low abundant cell types were well-predicted, such as plasmablasts, and thus the abundance is not the sole determinant of performance.

BLADE significantly outperformed CIBERSORTx in the estimation of gene expression profiles per cell type in both group-mode and high-resolution mode (**Fig. 6c-e**). For group-mode purification, CIBERSORTx reconstructed expression profiles per cell type with reasonable accuracy (>0.5 Pearson correlation except for pDC). The highest performance was achieved for classical monocytes. Here, performance of BLADE was near-perfect (**Fig. 6c**). In high-resolution mode, CIBERSORTx did not estimate expression levels of most genes, and essentially no genes were *in silico* purified for 11 cell types (**Fig. 6d**). Furthermore, even after filtering, the estimated gene expression profiles per cell type and per sample by CIBERSORTx are less accurate than those by BLADE (**Fig. 6e**). The performance of BLADE in high-resolution mode purification is consistently accurate (>0.7 Pearson correlation) across all 15 cell types. Cumulatively, Bayesian simultaneous

deconvolution and *in silico* purification by BLADE significantly outperformed CIBERSORTx in both estimating cellular fraction and especially in reconstructing gene expression profiles per cell type.

Discussion

One of the major challenges in deconvolution of bulk RNAseq data is adequate and yet efficient handling of gene expression variability without log-normalization. This difficulty causes either of the two critical shortcomings in most of the available deconvolution algorithms: 1) only a small number of cell types can be handled; and 2) the inadequate variability model (usually normal distribution) implicitly or explicitly assumed in the core algorithm, such as support vector regression and non-negative least-squares, implies inferior results. We showed that, the normal distribution often renders a biased fit (**Fig. 2b-c**). The inadequate noise model leads to suboptimal performance of deconvolution algorithms when there is a realistic level of gene expression variability (**Fig. 5**). Furthermore, purification of gene expression involves more variables to be estimated, for which an incorrect variability model can have substantial impact on the performance (**Figs. 5 and 6b-d**). Statistical inference of log-normal convolution models, which were shown to appropriately capture variability, however, is very challenging, as demonstrated by previous log-normal deconvolution methods that handle three cell types maximally [18].

BLADE solves this by using a novel hierarchical Bayesian model that simultaneously performs deconvolution and estimation of gene expression profiles per cell type. The log-normal convolution model efficiently accounts for variability in gene expression (**Fig. 3**) and also for prior knowledge of gene expression profiles per cell type derived from scRNA-seq data (**Fig. 4**). Notably, thanks to the unified probabilistic model used in BLADE, the prior knowledge contributes to both deconvolution and gene expression purification. This prior knowledge significantly reduces the search space of solutions for both tasks, which leads to enhanced accuracy and coverage, especially for gene expression purification. The efficient

variational inference of BLADE allowed to handle a large number of cell types (> 20 cell types) which was not possible by previous statistical approaches [17,18]. Furthermore, BLADE may be beneficial in handling cell types without a precise prior knowledge, for instance, cancer cells with highly variable gene expression profiles across the subject, unlike the non-malignant cells [24]. By integrating both signatures of mean and standard deviation, BLADE balances the contribution of genes with varying precision for deconvolution by prioritizing the genes with low variability (i.e., high precision) when estimating cell type fractions.

Enhanced *in silico* microdissection by BLADE opens up the possibility to molecularly characterize individual cell types in tissue based on the standard RNA-seq data. For instance, BLADE can be applied to estimate gene expression profiles of each cell type that makes up the tumor microenvironment (TME). This allows us to characterize pathway activity in each immune cell type, and possibly to recognize additional cell (sub-)types. Furthermore, BLADE can aid previously-established gene expression subtypes (e.g., PDAC [25,26]) by characterizing the subtypes with distinct TME profiles. The detailed profiling of the TME, in particular, immune TME profiles may lead to a clinically applicable biomarker strategy for immunotherapy based on the standard bulk gene expression profiling. In conclusion, BLADE is a novel tool that can significantly contribute to unravel cellular heterogeneity in complex biological systems.

Methods

PBMC single-cell RNA-seq data

Two public Peripheral Blood Mononuclear Cell (PBMC) CITE-seq (Cellular Indexing of Transcriptomes and Epitopes by Sequencing) datasets of healthy donors were downloaded from 10x Genomics (https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.2/5k_pbmc_protein_v3, https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_protein_v3). Genes and cells were filtered based on the following criterions: percentage of mitochondrial genes < 10% and number of

genes per cell between 200 and 4000. After the filtering, raw count data was normalized and scaled, using SCTransform, which performs normalization and variance stabilization using regularized negative binomial regression. Dimensionality reduction was done using principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-sne). Following that, k-nearest neighbors (knn) of each cell using 25 dimensions of PCA were determined. This knn graph was used to construct the Shared Nearest Neighbor (SNN) graph by calculating the neighborhood overlap (Jaccard index) between every cell and its 20 nearest neighbors. Cluster determination was done by SNN graph modularity optimization based on the Louvain algorithm with the resolution of 1. Cells were phenotyped separately in both datasets, using primarily cell surface markers and then gene expression levels in case of lack of usable cell surface markers (**Supplementary Fig. S1-2**). The two datasets were batch-corrected and integrated as described by Stuart T, et al[27]. Differentially expressed genes per cell type were identified using a Wilcoxon Rank sum test by taking a contrast between one cell type versus the rest.

A generic deconvolution method with known cellular composition

For a fair comparison of log-normal and negative binomial distribution for deconvolution, we developed a simple, generic maximum-likelihood based convolution model. Formally it is assumed that there are $i = 1, \ldots, I$ samples in which $t = 1, \ldots, T$ cell types jointly contribute to expression profiles of $j = 1, \ldots, J$ genes. For each sample i and gene j, a bulk expression level is given, indicated by y_{ij}. As in other deconvolution methods, two hidden variables were introduced that jointly makeup y_{ij}: 1) expression level of the gene per cell type, $x^t{}_{ij}$; and 2) cellular composition for each cell type t, $f^t{}_i$, where $\forall f^t{}_i \geq 0$ and $\sum_t f^t{}_i = 1$. An important strength of our method here is that it applies to any underlying parametric distribution for $x^t{}_{ij}$. y_{ij} is a (weighted) convolution:

$$y_{ij} = \sum_{t=1}^{T} f^{t}{}_{i} x^{t}{}_{ij} \quad (1)$$

which implies, with $\hat{x}^{t}{}_{ij} = f^{t}{}_{i} x^{t}{}_{ij}$,

$$g_{y_{ij}}(y) = \int_{u_1=0}^{y} \cdots \int_{u_T=0}^{y-\sum_{t=1}^{T-1} u_t} g_{\hat{x}^{1}{}_{ij}}(u_1) \cdots g_{\hat{x}^{T-1}{}_{ij}}(u_{T-1}) g_{\hat{x}^{T}{}_{ij}}(y - \sum_{t=1}^{T-1} u_t) du_1 \cdots du_T. \quad (2)$$

By assuming $x^{t}{}_{ij}$ follows log-normal distribution (i.e., $x^{t}{}_{ij} \sim LN(\mu^{t}{}_{j}, (\sigma^{t}{}_{j})^2)$) and thus $\hat{x}^{t}{}_{ij} \sim LN(\mu^{t}{}_{j} + \log f^{t}{}_{i}, (\sigma^{t}{}_{j})^2)$, y_{ij} is a convolution of T log-normal random variables. The interest lies in estimating parameters $\theta_j = (\mu^{t}{}_{j}, \sigma^{t}{}_{j})$ by maximum likelihood.

While numerical evaluation of (2) may still be efficient for $T = 2$ [28], however, the extension to $T > 2$ is not straightforward to a $T - 1$ dimensional integral. To this end, the log-normal density $g_t = g_{\hat{x}^{t}{}_{ij}}$ is approximated by a probability generating function (PGF). See **Supplementary Note 1** for the details of PGF approximation. The PGF-based approximation of g_t showed higher accuracy than an alternative approximation method, Fenton-Wilkinson (FW) approximation[29], which was also included as a benchmark (See **Supplementary Methods** and **Supplementary Fig. S3**).

Comparison of MLE for LN and NB based on the generic deconvolution technique

The aforementioned generic deconvolution was used to evaluate LN and NB for deconvolution. For this, two RNA-seq data sets are retrieved from The Cancer Genome Atlas (https://tcga-data.nci.nih.gov/tcga/) using TCGAbiolinks[30]. We considered all complete samples from the following tumor types: Mesothelioma (MESO[31], n = 84; and Sarcoma (SARC[32], n = 256. Data was preprocessed as described previously in Rauschenberger et al. [33]. The comparison procedure for LN and NB distributions is:

1. Apply a non-statistical method, EPIC[19], to estimate cell type fractions for bulk RNA-seq data using cell type specific reference signatures. It has shown that EPIC provides an

reliable estimate of cellular fractions of $T = 8$ cell types[34], and it provides absolute fractions that add up to 1.

2. Fix the cellular fractions and fit generic deconvolution models with $T = 8$ LN or NB components using maximum likelihood.

3. Compare the maximum likelihood values of the LN and NB models for J genes.

The above procedure was done for 200 randomly selected genes with mean count per million larger or equal to 5 to exclude lowly expressed genes. Note that the comparison of the maximum likelihood values is fair, because the number of parameters used in the LN and NB components is the same, $2T = 16$ per gene.

Hierarchical Bayesian model for convolution of log-normal variables (BLADE)

A novel Bayesian Log-normal Deconvolution model, BLADE, is introduced to efficiently perform log-normal convolution, while accounting for the prior knowledge of per cell-type gene expression profiles (see Overview at **Fig. 4a**). Formally, we assume $y_{ij} = \sum_t f^t{}_i x^t{}_{ij} + \epsilon_{ij}$, where ϵ_{ij} is a log-normal error with mean parameter 0 and variance parameter γ_j. Then, $x^t{}_{ij}$ follows a log-normal distribution: $x^t{}_{ij} \sim LN(\mu^t{}_j, \frac{1}{\lambda_j})$, where $\mu^t{}_j$ and $\lambda^t{}_j$ are expected value and precision in log-scale. Note that the parameters $\mu^t{}_j$ and $\lambda^t{}_j$ are shared across the samples. To incorporate prior knowledge on gene expression profiles per cell type, a hierarchical Bayesian approach was taken: $\mu^t{}_j$ and $\lambda^t{}_j$ are endowed with normal-gamma priors with hyperparameters $\mu^t{}_{0j}$, $\kappa^t{}_{0j}$, $\alpha^t{}_{0j}$, and $\beta^t{}_{0j}$: $(\mu^t{}_j, \lambda^t{}_j) \sim NG(\mu^t{}_{0j}, \kappa^t{}_{0j}, \alpha^t{}_{0j}, \beta^t{}_{0j})$. Note that the normal-gamma distribution is a conjugate prior of log-normal distribution, based on which marginal distribution of $x^t{}_{ij}$ given the hyperparameters $\mu^t{}_{0j}$, $\kappa^t{}_{0j}$, $\alpha^t{}_{0j}$, and $\beta^t{}_{0j}$ is analytically tractable. The other hidden variable, $f^t{}_i$, was endowed with Dirichlet distribution: $(f^t{}_1, \ldots, f^t{}_I) \sim D(\alpha^t{}_1, \ldots, \alpha^t{}_I)$.

For the inference, a collapsed variational inference was employed to handle analytically intractable posterior distribution of hidden variables given observed variables[35]. In the framework, the random variables with conjugate prior distribution, which are ${\mu^t}_j$ and ${\lambda^t}_j$, were integrated out, which allows us to find a fully Bayesian estimation of ${x^t}_{ij}$ instead of estimation of the single most probable ${\mu^t}_j$ and ${\lambda^t}_j$ [35]. By defining the variational distribution for the hidden variables, ${x^t}_{ij}$ and ${f_i}^t$, the objective function is to minimize the dissimilarity between the variational distribution and probability distribution, measured by Kullback-Leibler divergence (see **Supplementary Note 2** for the detailed derivation). The minimization was done by the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm available in SciPy Python library with the constraints of ${f_i}^t \geq 0$ and $\sum_t {f_i}^t = 1$. Numba-compiled objective function and gradients were used for the acceleration.

Selection of hyperparameters based on the empirical-Bayes framework

BLADE has multiple hyperparameters for the hidden variables ${x^t}_{ij}$ and ${f_i}^t$, and also for observed variable y_{ij}. For ${f_i}^t$, a hyperparameter ${\alpha^t}_i$ for Dirichlet distribution is required, for which we chose one value across the different ts since we do not have prior information on cellular composition. For y_{ij}, we need to specify a precision of each gene, γ_j, which we chose $\frac{s}{\mathrm{var}(\log y_{ij})}$, where s and $\mathrm{var}(\log y_{ij})$ are a user-defined scale factor and a variance in log-scale measured per gene, respectively. For hyperparameters of ${x^t}_{ij}$, ${\mu^t}_{0j}$, ${\kappa^t}_{0j}$, ${\alpha^t}_{0j}$, and ${\beta^t}_{0j}$, we incorporated prior knowledge of gene expression profiles per cell type obtained from the scRNA-seq data. Given log-normal likelihood and normal-gamma priors, average expression level and standard deviation of ${x^t}_{ij}$ are: $E(\log {x^t}_{ij}) = {\mu^t}_{0j}$ and $\mathrm{var}(\log {x^t}_{ij}) = \frac{{\beta^t}_{0j}}{{\alpha^t}_{0j}}$, respectively. To make use of the prior knowledge, we obtained the sample estimates of $E(\log {x^t}_{ij})$ and $\mathrm{var}(\log {x^t}_{ij})$ from the scRNA-seq data,

denoted by $\mu^t{}_j$ and $(\sigma^t{}_j)^2$. Then, we assigned $\mu^t{}_{0j} = \mu^t{}_j$ whereas $\alpha^t{}_{0j}$ is set by users followed by deriving: $\beta^t{}_{0j} = \alpha^t{}_{0j}(\sigma^t{}_j)^2$. Here, $\alpha^t{}_{0j}$ allows to adapt to how much information the single cell data carries for the bulk RNA-seq data. The other hyperparameter $\kappa^t{}_{0j}$ is also user-defined, which serve as a scale factor for variance of $\mu^t{}_j$ (see also **Supplementary Note 2**).

An empirical Bayes approach was employed to select the best set of user-defined parameters[36]. For each configuration of parameters, a maximum likelihood estimate of variational parameters is obtained using a subset of samples. Then, the hyperparameter configuration with the highest likelihood is selected, followed by performing deconvolution using the entire data set. Only a subset of samples is used in the empirical Bayes step, not only to gain computational efficiency but also to avoid overfitting. Throughout the manuscript, we considered a total of 90 different parameter configurations that cover all possible combinations of: $\alpha^t{}_i \in \{1,10\}$, $\alpha^t{}_{0j} \in \{0.1, 0.5, 1, 5, 10\}$, $\kappa^t{}_{0j} \in \{1, 0.5, 0.1\}$, and $s \in \{1, 0.3, 0.5\}$.

Construction of the simulation data with a controlled noise level

We constructed simulation data sets of bulk gene expression profiles with known cellular fraction, gene expression profiles per cell type, and a diverse number of cell types and samples. To this end, given a number of cell types and genes, we first randomly sample an expected gene expression level $\mu^t{}_j$ for gene j and cell type t from a normal distribution with 0 mean and standard deviation of 1.5: $\mu^t{}_j \sim N(0, 2)$. Then, we sample gene expression levels per sample and per cell type, $x^t{}_{ij}$ from a log-normal distribution with mean $\mu^t{}_j$ and standard deviation of σ ($x^t{}_{ij} \sim LN(\mu^t{}_j, \sigma)$), where σ is the parameter to control the variability in gene expression per cell type of each simulation data set. Fraction of cell types are sampled from a Dirichlet distribution with uninformative prior: $f_i{}^t \sim D(\forall_i \alpha^t{}_i)$, where

$\alpha^t{}_i = 1$. Then, the bulk gene expression profiles are generated by: $y_{ij} = \sum_t f_i{}^t x_{ij}{}^t$. We constructed in total of 700 training data sets with the following settings: 1) number of samples = [5,10,20,50,100]; 2) number of genes = [100,200,500,1000]; 3) number of cell types = [2,3,5,10,20]; and 4) level of variability in gene expression profiles per cell type: σ=[0.1,0.2,0.5,0.75,1,1.25,1.5].

Construction of PBMC stimulation data

To construct realistic simulation data, 20 bulk gene expression data sets were generated by randomly sampling and merging a subset of 10,403 cells from the two PBMC scRNA-seq datasets. For each sample, the cellular fraction was first sampled from a Dirichlet distribution. The actual fractions of the 15 cell types were used as the parameter of the Dirichlet distribution so that the sampled fraction is similar to the total fraction. The fraction was then converted into the count of each cell type, with the following constraints: 1) the total number of cells is 100, and 2) the minimum number of cells per type is one. Then, the given number of cells were sampled with replacement, followed by obtaining the raw counts per cell type as the cumulative sum of raw counts of the sampled cells. Up to three distinct cells per type were allowed to be sampled since otherwise, gene expression variability was over-stabilized due to the averaging. Finally, the simulated bulk raw counts were obtained by taking the cumulative sum of the raw counts per cell type among 15 cell types. The bulk gene expression data was log-normalized using the Seurat package[27].

Systematic evaluation of BLADE and comparison against baseline methods

The original docker image of CIBERSORTx and NNLS, were obtained from https://cibersortx.stanford.edu/ and from the SciPy Python library, respectively. For all three methods, the same signatures of average gene expression profiles per cell type were used, while BLADE also used standard deviations of each gene. For the simulation data sets with

the controlled gene expression variability level, true mean $\mu^t{}_j$ and level of variability per cell type of all genes were retrieved. For the PBMC bulk transcriptome data, average and standard deviations of the union of differentially expressed genes of 15 cell types (in total 1,007 genes) were obtained from the PBMC scRNAseq data. For evaluation of the deconvolution performance, the Pearson correlation was measured between the predicted and true abundance of each cell type across the samples. For evaluation of purification, the Pearson correlation was measured between true and estimated gene expression profiles per cell type for group mode. For the high-resolution mode, the Pearson correlation was measured per sample. The performance evaluation for purification was done only for CIBERSORTx and BLADE as NNLS only estimate cellular fractions.

Code availability

BLADE python software along with a user-friendly demo is available and maintained at https://github.com/tgac-vumc/BLADE

Acknowledgments

This project was supported by stichting Cancer Center Amsterdam (CCA2019-9-62).

Author contributions

Y.K. and M.W. conceived the ideas, and designed the algorithm. Y.K. and B.A.B. developed the python software. B.A.B. and S.D.A. analyzed PBMC Cite-seq data. B.A.B., S.D.A. and J.G.V. classified immune cell types in the CITE-seq data. Biological interpretation of the outcome is done by S.D.A., J.O., A.F.S., J.V., F.D., H.L. B.Y., and J.G.V. Evaluation of the algorithm performance is designed and performed by Y.K., and B.A.B. All authors discussed the results and contributed to the writing.

References

1. Angelova, M. *et al.* Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.* **16**, 1–17 (2015).
2. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (80-.).* **352**, 189–196 (2016).
3. Zheng, Z. *et al.* Single-cell transcriptomic analysis. *Compr. Physiol.* **10**, 767–783 (2020).
4. Pottier, C. *et al.* The importance of the tumor microenvironment in the therapeutic management of cancer. *Expert Rev. Anticancer Ther.* **15**, 943–954 (2015).
5. Kumar, M. P. *et al.* Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. *Cell Rep.* **25**, 1458-1468.e4 (2018).
6. Xiao, Z., Dai, Z. & Locasale, J. W. Metabolic landscape of the tumor microenvironment at single cell resolution. *Nat. Commun.* **10**, 1–12 (2019).
7. Sathe, A. *et al.* Single-Cell Genomic Characterization Reveals the Cellular Reprogramming of the Gastric Tumor Microenvironment. *Clin. Cancer Res.* **26**, 2640–2653 (2020).
8. Xiao, Z., Locasale, J. W. & Dai, Z. Metabolism in the tumor microenvironment: insights from single-cell analysis. *Oncoimmunology* **9**, (2020).
9. Lee, H. W. *et al.* Single-cell RNA sequencing reveals the tumor microenvironment and facilitates strategic choices to circumvent treatment failure in a chemorefractory bladder cancer patient. *Genome Med.* **12**, 1–21 (2020).
10. González-Silva, L., Quevedo, L. & Varela, I. Tumor Functional Heterogeneity Unraveled by scRNA-seq Technologies. *Trends in Cancer* **6**, 13–19 (2020).
11. Angerer, P. *et al.* Single cells make big data: New challenges and opportunities in transcriptomics. *Curr. Opin. Syst. Biol.* **4**, 85–91 (2017).
12. Lähnemann, D. *et al. Eleven grand challenges in single-cell data science*. *Genome*

Biology vol. 21 (Genome Biology, 2020).

13. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
14. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).
15. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, (2019).
16. Zhong, Y. & Liu, Z. Gene expression deconvolution in linear space. *Nat. Methods* **9**, 8–9 (2012).
17. Wang, L. *et al.* A reference profile-free deconvolution method to infer cancer cell-intrinsic subtypes and tumor-type-specific stromal profiles. *Genome Med.* **12**, 1–22 (2020).
18. Wang, Z. *et al.* Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. *iScience* **9**, 451–460 (2018).
19. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife* **6**, 1–25 (2017).
20. Monaco, G. *et al.* RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Rep.* **26**, 1627-1640.e7 (2019).
21. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
22. Lam, S. K., Pitrou, A. & Seibert, S. Numba: a LLVM-based Python JIT compiler. *Proc. Second Work. LLVM Compil. Infrastruct. HPC - LLVM '15* 1–6 (2015) doi:10.1145/2833157.2833162.
23. Li, T. *et al.* TIMER: A web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res.* **77**, e108–e110 (2017).

24. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e24 (2017).

25. Dijk, F. *et al.* Unsupervised class discovery in pancreatic ductal adenocarcinoma reveals cell-intrinsic mesenchymal features and high concordance between existing classification systems. *Sci. Rep.* **10**, 1–12 (2020).

26. Bailey, P. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**, 47–52 (2016).

27. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

28. Ahn, J. *et al.* DeMix: Deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics* **29**, 1865–1871 (2013).

29. Fenton, L. F. The Sum of Log-Normal Probability Distributions in Scatter Transmission Systems. *IRE Trans. Commun. Syst.* **8**, 57–67 (1960).

30. Colaprico, A. *et al.* TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71 (2016).

31. Hmeljak, J. *et al.* Integrative molecular characterization of malignant pleural mesothelioma. *Cancer Discov.* **8**, 1549–1565 (2018).

32. Genomic, I., Of, C., Soft, A. & Sarcomas, T. Comprehensive and Integrated Genomic CHARACTERIZATION OF ADULT SOFT TISSUE SARCOMAS. **171**, 950–965 (2018).

33. Rauschenberger, A., Ciocănea-Teodorescu, I., Jonker, M. A., Menezes, R. X. & van de Wiel, M. A. Sparse classification with paired covariates. *Adv. Data Anal. Classif.* **14**, 571–588 (2020).

34. Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).

35. Teh, Y. W., Newman, D. & Welling, M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Adv. Neural Inf. Process. Syst.* 1353–1360 (2007) doi:10.7551/mitpress/7503.003.0174.

36. Carlin, B. P. & Louis, T. A. Empirical Bayes: Past, Present and Future. *J. Am. Stat.*

Assoc. **95**, 1286–1289 (2000).

FIGURES LEGENDS

Fig. 1. Overview of single-cell RNA-seq data from two PBMC samples. a. t-SNE plots show gene-expression variability of two single-cell PBMC RNA-seq data, on left and right, respectively. Each unique cell type is denoted by color. **b-c.** Comparison of gene expression variability (standard deviation per gene; y-axis of box plots) in log-scale (b) and linear-scale (c) of both datasets (x-axis). Variability is measured per gene and per cell type, and they are split by differentially expressed genes (red) and non-differentially expressed genes (DEG; blue). **d.** Comparison of within-sample (x-axis) and between-sample variability (y-axis) in gene expression levels per cell type. Standard deviation is measured for each gene and cell type first separately in two PBMC single-cell datasets followed by taking average (x-axis), then also in merged PBMC data set (y-axis). Only the 9 cell types commonly detected in two data sets were used in the analysis.

Fig. 2. Comparison of normal, negative binomial, and log-normal distribution in fitting linear-scale gene expression data. For each gene and cell type, maximum-likelihood parameter estimation is done for each of the three distribution assumptions. **a**. A bar chart of average log-likelihood of the three types of distribution fitted to PBMC single-cell RNA-seq data. The genes are split by DEGs (red) and non-DEGs (blue). **b.** Density plots for raw-counts (red) and optimized log-normal (green), normal (blue), and negative binomial distribution (purple) for two example genes, HLA-DRA (top) and CD74 (bottom). **c**. Pairwise comparison of per-gene log-likelihood of log-normal distribution (top (y-axis) and middle (y-axis)) and that of normal (top (x-axis) and bottom (y-axis)) and negative binomial distribution (middle (x-axis) and bottom (x-axis)). The genes are split into non-DEGs (left) and DEGs (right).

Fig. 3. Comparison of negative binomial and log-normal components in generic deconvolution of gene expression data from TCGA. Maximum log-likelihood values of each gene for log-normal (y-axis) and negative binomial (bottom) convolutions of T=8 cell types, applied to TCGA-MESO (left) and TCGA-SARC (right) data. Cellular fractions are pre-estimated by EPIC.

Fig. 4. BLADE workflow. a. To construct a prior knowledge for BLADE, we used CITE-seq data that contains gene expression and cell surface marker profiles of each cell. Cells are then subject to phenotyping, clustering, and differential gene expression analysis. Then, for each cell type, we retrieved average expression profiles (red cross and top heat map) and standard deviation per gene as the variability (blue circle and bottom heatmap). This prior knowledge is then used in the hierarchical Bayesian model (bottom right) to deconvolute bulk transcriptome profile. **b.** A Bayesian network of BLADE represents random variables, observed and hidden variables respectively in blue and gray nodes, and their dependency associations (arrows). The hyperparameters (${\mu^t}_{0j}$, ${\kappa^t}_{0j}$, ${\alpha^t}_{0j}$, and ${\beta^t}_{0j}$ on the left) and bulk gene expression data (y_{ij} on the right) are observed, where the hidden variables in the middle (${\mu^t}_j$, ${\lambda_j}^t$, ${f_i}^t$ and ${x_{ij}}^t$) are inferred. Among the hidden variables, ${f^t}_i$ and ${x_{ij}}^t$ respectively represent the fraction of cell type t and purified gene expression level of gene j in sample i.

Fig. 5. Performance evaluation BLADE in simulation data with diverse settings. a. Performances (Pearson correlation; y-axis) of BLADE (orange), CIBERSORTx (blue), and NNLS (green) in predicting cellular fraction of simulation data with diverse variability level (standard deviation of 0.1-1.5; x-axis), number of cells (2-20 cell types; rows), and number of genes (100-1000 genes; columns). **b-c.** Performances (Pearson correlation; y-axis) of BLADE (orange) and CIBERSORTx (blue) in predicting gene expression profiles per cell type for all samples jointly (group-mode; **b**) and for each sample separately (high-resolution mode; **c**) using the same simulation data. **d-e.** Fractions of genes with the inferred gene

expression profiles per cell type by CIBERSORTx in group-mode (**d**) and high-resolution mode (**e**). x- and y-axis represents the number of cell types and samples in the simulation data, respectively.

Fig. 6. Performance evaluation of BLADE in simulated PBMC bulk RNA-seq data. a. t-SNE plot showing variability of the cell populations in the simulated bulk RNA-seq data. **b-c**. Radar charts represent performance in Pearson correlation of BLADE (orange), CIBERSORTx (blue), and NNLS (green) for estimation of cellular fractions (**b**) and for group-mode purification (**c**). **d**. Fraction of genes *in* silico purified in group-mode (blue) and high-resolution-mode (red) by CIBERSORTx. **e**. Performance (Pearson correlation; y-axis) of BLADE (orange) and CIBERSORTx (blue) in estimating gene expression profiles per cell type and per sample.

Figures

Figure 1

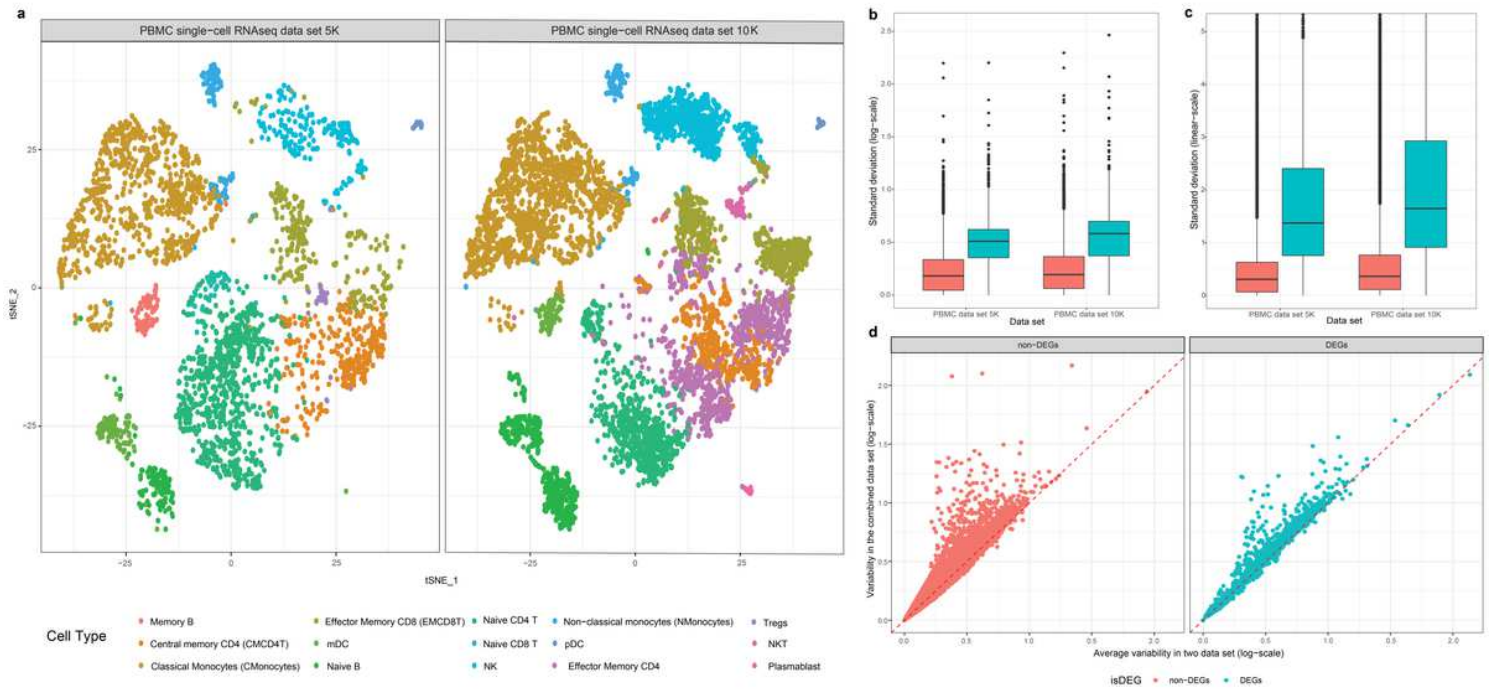

Figure 1

Overview of single-cell RNA-seq data from two PBMC samples. a. t-SNE plots show gene-expression variability of two single-cell PBMC RNA-seq data, on left and right, respectively. Each unique cell type is denoted by color. b-c. Comparison of gene expression variability (standard deviation per gene; y-axis of box plots) in log-scale (b) and linear-scale (c) of both datasets (x-axis). Variability is measured per gene and per cell type, and they are split by differentially expressed genes (red) and non-differentially expressed genes (DEG; blue). d. Comparison of within-sample (x-axis) and between-sample variability (y-axis) in gene expression levels per cell type. Standard deviation is measured for each gene and cell type first separately in two PBMC single-cell datasets followed by taking average (x-axis), then also in merged PBMC data set (y-axis). Only the 9 cell types commonly detected in two data sets were used in the analysis.

Figure 2

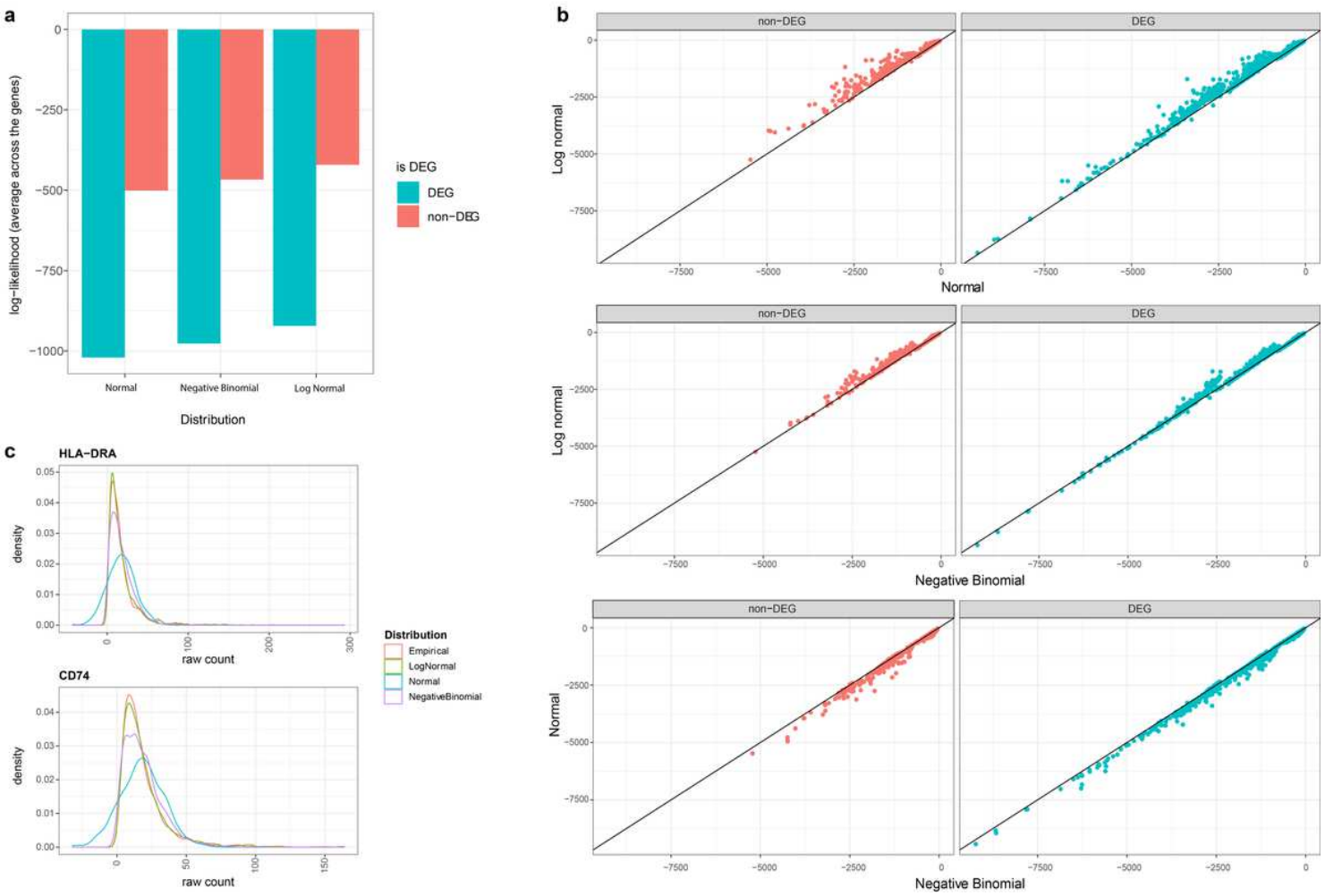

Figure 2

Comparison of normal, negative binomial, and log-normal distribution in fitting linear-scale gene expression data. For each gene and cell type, maximum-likelihood parameter estimation is done for each of the three distribution assumptions. a. A bar chart of average log-likelihood of the three types of distribution fitted to PBMC single-cell RNA-seq data. The genes are split by DEGs (red) and non-DEGs (blue). b. Density plots for rawcounts (red) and optimized log-normal (green), normal (blue), and negative binomial distribution (purple) for two example genes, HLA-DRA (top) and CD74 (bottom). c. Pairwise comparison of per-gene log-likelihood of log-normal distribution (top (y-axis) and middle (yaxis)) and that of normal (top (x-axis) and bottom (y-axis)) and negative binomial distribution (middle (x-axis) and bottom (x-axis)). The genes are split into non-DEGs (left) and DEGs (right).

Figure 3

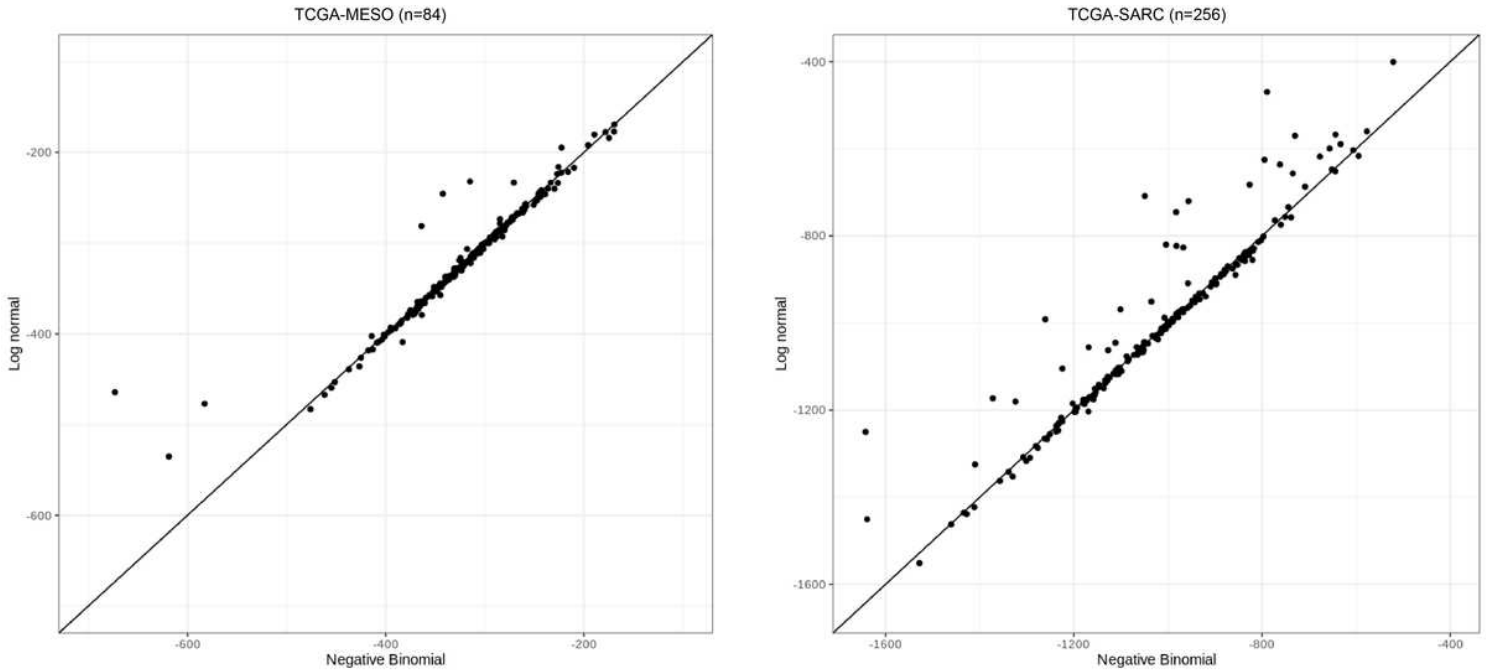

Figure 3

Comparison of negative binomial and log-normal components in generic deconvolution of gene expression data from TCGA. Maximum log-likelihood values of each gene for log-normal (y-axis) and negative binomial (bottom) convolutions of T=8 cell types, applied to TCGA-MESO (left) and TCGA-SARC (right) data. Cellular fractions are preestimated by EPIC.

Figure 4

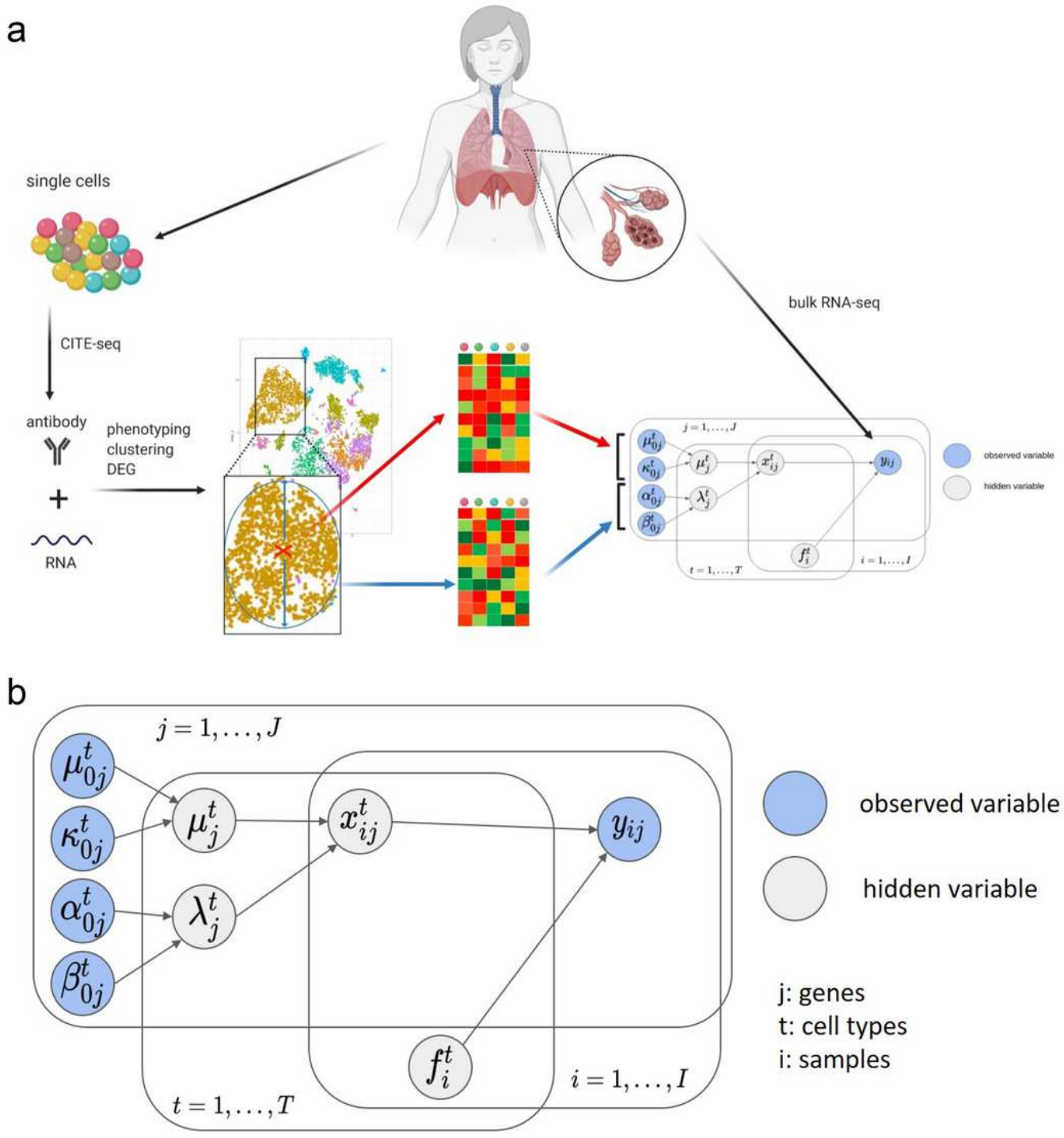

Figure 4

BLADE workflow. a. To construct a prior knowledge for BLADE, we used CITE-seq data that contains gene expression and cell surface marker profiles of each cell. Cells are then subject to phenotyping, clustering, and differential gene expression analysis. Then, for each cell type, we retrieved average expression profiles (red cross and top heat map) and standard deviation per gene as the variability (blue circle and bottom heatmap). This prior knowledge is then used in the hierarchical Bayesian model (bottom right) to

deconvolute bulk transcriptome profile. b. A Bayesian network of BLADE represents random variables, observed and hidden variables respectively in blue and gray nodes, and their dependency associations (arrows). The hyperparameters (μt0j, κt0j, αt0j, and βt0j on the left) and bulk gene expression data (yijon the right) are observed, where the hidden variables in the middle (μtj, λjt, fit, and xijt) are inferred. Among the hidden variables, fti and xijt respectively represent the fraction of cell type t and purified gene expression level of gene j in sample i.

Figure 5

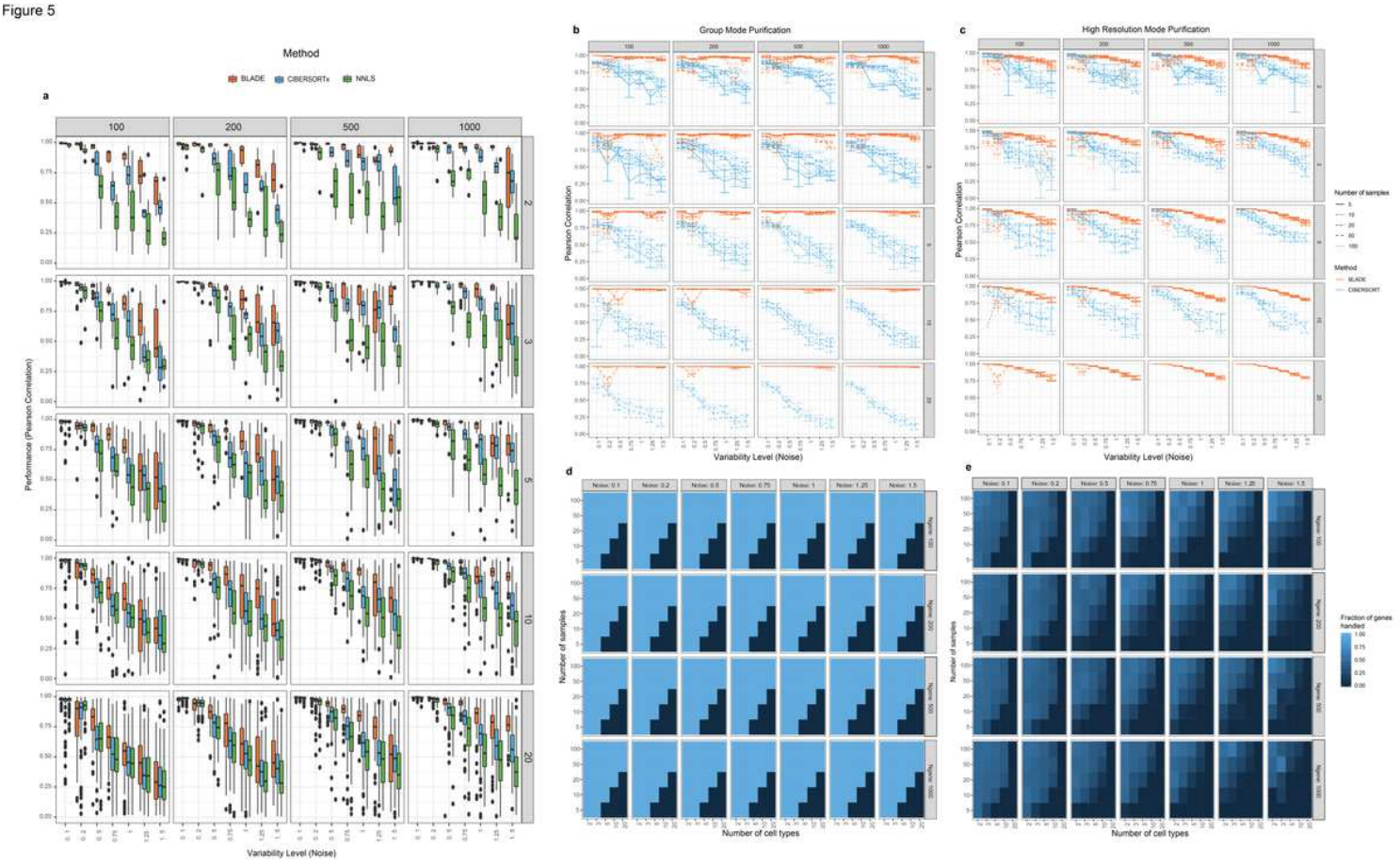

Figure 5

Performance evaluation BLADE in simulation data with diverse settings. a. Performances (Pearson correlation; y-axis) of BLADE (orange), CIBERSORTx (blue), and NNLS (green) in predicting cellular fraction of simulation data with diverse variability level (standard deviation of 0.1-1.5; x-axis), number of cells (2-20 cell types; rows), and number of genes (100-1000 genes; columns). b-c. Performances (Pearson correlation; y-axis) of BLADE (orange) and CIBERSORTx (blue) in predicting gene expression profiles per cell type for all samples jointly (group-mode; b) and for each sample separately (high-resolution mode; c) using the same simulation data. d-e. Fractions of genes with the inferred gene expression profiles per cell type by CIBERSORTx in group-mode (d) and high-resolution mode (e). x- and y-axis represents the number of cell types and samples in the simulation data, respectively.

Figure 6

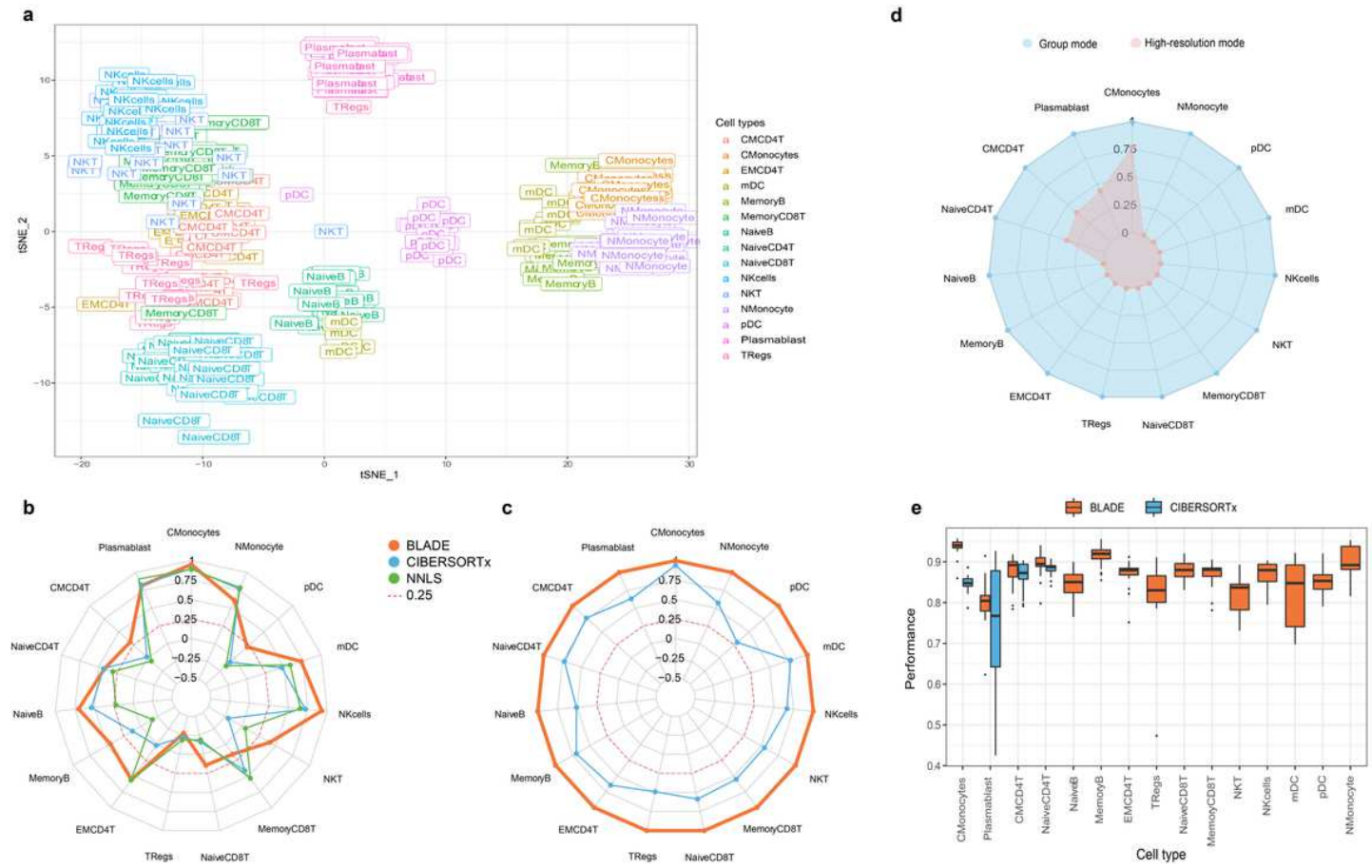

Figure 6

Performance evaluation of BLADE in simulated PBMC bulk RNA-seq data. a. t- SNE plot showing variability of the cell populations in the simulated bulk RNA-seq data. b-c. Radar charts represent performance in Pearson correlation of BLADE (orange), CIBERSORTx (blue), and NNLS (green) for estimation of cellular fractions (b) and for groupmode purification (c). d. Fraction of genes in silico purified in group-mode (blue) and highresolution- mode (red) by CIBERSORTx. e. Performance (Pearson correlation; y-axis) of BLADE (orange) and CIBERSORTx (blue) in estimating gene expression profiles per cell type and per sample.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.pdf
- SupplementaryNotes.pdf