GLDADec: marker-gene guided LDA modelling for bulk gene expression deconvolution

Iori Azuma[*, 1], *Tadahaya Mizuno*[*, †, 2], *and Hiroyuki Kusuhara*[‡, 1]

AUTHOR ADDRESS

[1]Graduate School of Pharmaceutical Sciences, the University of Tokyo, Bunkyo-ku, Tokyo, 113-0033, Japan

[2]Graduate School of Pharmaceutical Sciences, the University of Tokyo, Bunkyo-ku, Tokyo, 113-0033, Japan, Electronic address: tadahaya@gmail.com

† Corresponding author: Tadahaya Mizuno, tadahaya@gmail.com; Graduate School of Pharmaceutical Sciences, the University of Tokyo, Bunkyo-ku, Tokyo, 113-0033, Japan; Tel: +81-3-5841-4771

*These authors contributed equally.

Abstract

Inferring cell type proportions from bulk transcriptome data is crucial in immunology and oncology. Here, we introduce GLDADec (Guided LDA Deconvolution), a bulk deconvolution method that guides topics using cell type-specific marker gene names to estimate topic distributions for each sample. Through benchmarking using blood-derived datasets, we demonstrate its high estimation performance and robustness. Moreover, we apply GLDADec to heterogeneous tissue bulk data and perform comprehensive cell type analysis in a data-driven manner. We show that GLDADec outperforms existing methods in estimation performance and evaluate its biological interpretability by examining enrichment of biological processes for topics. Finally, we apply GLDADec to TCGA tumor samples, enabling subtype stratification and survival analysis based on estimated cell type proportions, thus proving its practical utility in clinical settings. This approach, utilizing marker gene names as partial prior information, can be applied to various scenarios for bulk data deconvolution. GLDADec is an available as an open-source Python package at https://github.com/mizuno-group/GLDADec.

Keywords: deconvolution, marker gene name, semi-supervised learning, perturbation, TCGA, LDA

Introduction

Quantification of the proportion of cell types in a tissue sample and understanding the contribution of individual cell types to the physiological state, such as immune responses associated with perturbation or evaluation of cancer tumor samples with cell proliferation, is of utmost importance [1,2]. Flow cytometry, a typical experimental approach for quantifying compositional proportions of cell types, is widely used in molecular biology and immunology. However, its application is limited to fresh organs, human tissue specimens are challenging to analyze, and there is limited knowledge on the aggregation of flow cytometry data, making it difficult to utilize legacy data.

The advancement of high-throughput sequencing technology has led to an abundance of stored transcriptome data [3,4]. The bulk transcriptome measures the accumulation of gene expression levels derived from various cell types and can be applied for extensive analyses using several well-established databases [3–5]. Although databases for single-cell sequencing technology, a recent innovation, have also been developed, its high cost and sparce nature make it challenging to conduct large-scale data analyses [6]. Hence, it is useful to establish a method to estimate the proportion of constitutive cells from the bulk transcriptome.

Deconvolution is a computational method that can be used to estimate the proportion of immune cells in a sample using transcriptome data. In recent years, several deconvolution methods have been proposed to infer cell type proportions from bulk expression data [7–15]. These methods can be categorized into two main groups: reference-free methods and reference-based methods [16,17] Reference-free methods estimate cell type proportions solely based on the samples to be analyzed, making them less sensitive to external information that may cause confounding factors. This approach is reasonable for tissue data where the constituent cell types are not clearly defined, as the number of components can be estimated using likelihood and other factors [11,12]. However, identifying the estimated components and their corresponding cells is challenging, and the downstream task is less interpretable. In contrast, reference-based methods utilize cell type-specific gene expression profiles, called references, as prior information. Although there have been some notable successes [13–15], the performance of these methods depends on the quality of the reference data and the batch-to-batch differences between the data to be analyzed. As a result, reference-based methods are only applicable in specific scenarios where the major cellular components are well-defined and appropriate reference data are available. While they are effective for simulated datasets or samples with well-defined constituent cell types, such as blood, they may underestimate the impact of gene expression profiles derived from cells that are not assumed as the reference [18].

Latent Dirichlet Allocation (LDA) has been developed in the context of natural language processing, and it is widely used in various domains, such as text embedding or semantic extraction from documents [19,20]. The LDA model is designed to identify the topics that make up the content of documents, which is analogous to deconvolution, a method that extracts cell type-specific information from bulk transcriptome data. However, since LDA is an unsupervised learning method, it is classified as reference-free when simply applied to deconvolution, which presents a challenge in terms of cell identifiability. To address this issue, several approaches have been developed by incorporating prior information into the LDA algorithm and extending it to semi-supervised learning [12,21]. Although these ideas are reasonable, they rely on expression levels obtained from pure cell lines or single-cell RNA-Seq as prior information, which is expensive to obtain based on the tissues and species to be analyzed.

Here, we propose a novel Guided LDA Deconvolution (GLDADec) method, which utilizes marker gene names as partial prior information to estimate cell type proportions, thereby overcoming the challenges of conventional reference-based and reference-free methods simultaneously. This method employs a semi-supervised learning algorithm that combines cell-type marker genes with additional factors that may influence gene expression profiles to achieve a robust estimation of cell type proportions. Moreover, an ensemble strategy is used to aggregate the outputs and achieve a more accurate estimation. We benchmarked GLDADec against existing methods using blood-derived samples with well-defined constituent cells, and it consistently outperformed the existing methods for multiple datasets. We also applied GLDADec to liver bulk RNA-Seq data from drug-induced liver injury models of mice and rats, demonstrating its usefulness for tissue data analysis. Our model, which considers additional topics, reflects the biological processes underlying the tissue and provides a robust estimation of the guided target cell types. Additionally, by collecting marker gene names in a data-driven manner and using them as prior information on comprehensive cell types, it is possible to estimate a wider range of cell types that are inaccessible by conventional methods. As a further demonstration, we applied GLDADec to human tumor samples, revealing new insights into cancer subtyping and clinical prognostic stratification. GLDADec is available as an open-source package at https://github.com/mizuno-group/GLDADec.

Materials and Methods

Overview of GLDADec

Latent Dirichlet Allocation (LDA) algorithm for deconvolution

It is assumed that the sample-topic distribution $\boldsymbol{\theta}_m$ and topic-gene distribution $\boldsymbol{\varphi}_k$ are generated from the

Dirichlet distribution as follows:

$$\boldsymbol{\theta}_m \sim \mathrm{Dir}(\boldsymbol{\alpha}) \qquad (m = 1, \cdots, M) \tag{1.1}$$

$$\boldsymbol{\varphi}_k \sim \mathrm{Dir}(\boldsymbol{\eta}) \qquad (k = 1, \cdots, K), \tag{1.2}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ are hyperparameters. The latent topic $\boldsymbol{Z}_{m,n}$ and gene $\boldsymbol{G}_{m,n}$ are derived from multinomial distributions $\boldsymbol{Z}_{m,n} \sim \mathrm{Multi}(\boldsymbol{\theta}_m)$ and $\boldsymbol{G}_{m,n} \sim \mathrm{Multi}(\boldsymbol{\varphi}_{Z_{m,n}})$, respectively. The joint probability modeled by the graphical model (Figure 1) is given by:

$$P(\boldsymbol{G}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi} | \boldsymbol{\alpha}, \boldsymbol{\eta}) = \prod_m P(\boldsymbol{\theta}_m | \boldsymbol{\alpha}) \prod_k P(\boldsymbol{\varphi}_k | \boldsymbol{\eta}) \prod_n P(\boldsymbol{Z}_{m,i} | \boldsymbol{\theta}_m) P(\boldsymbol{G}_{m,n} | \boldsymbol{\varphi}_k, \boldsymbol{Z}_{m,n}). \tag{1.3}$$

With respect to the observed gene expression levels, the probability distribution that generated them is approximated through collapsed Gibbs sampling [22]. Specifically, the posterior distribution is expressed as follows:

$$\begin{aligned} & p(Z_{m,n} = k | G_{m,n} = v, \boldsymbol{G}^{\backslash m,n}, \boldsymbol{Z}^{\backslash m,n}, \boldsymbol{\alpha}, \boldsymbol{\eta}) \\ & \propto p(Z_{m,n} = k, G_{m,n} = v, \boldsymbol{G}^{\backslash m,n}, \boldsymbol{Z}^{\backslash m,n} \mid \boldsymbol{\alpha}, \boldsymbol{\eta}) \\ & \propto p(G_{m,n} = v | Z_{m,n} = k, \boldsymbol{G}^{\backslash m,n}, \boldsymbol{Z}^{\backslash m,n}, \boldsymbol{\alpha}, \boldsymbol{\eta})\, p(Z_{m,n} = k | \boldsymbol{G}^{\backslash m,n}, \boldsymbol{Z}^{\backslash m,n}, \boldsymbol{\alpha}, \boldsymbol{\eta}) \end{aligned} \tag{1.4}$$

Note that $\boldsymbol{G}^{\backslash m,n}$ is the distribution without counting $G_{m,n} = v$, and $\boldsymbol{Z}^{\backslash m,n}$ is without counting $Z_{m,n} = k$. For algorithmic convenience, we integrate out θ_m:

$$\int p(Z_{m,n} = k | \boldsymbol{\theta}_m)\, p(\boldsymbol{\theta}_m | \boldsymbol{Z}^{\backslash m,n}, \alpha) d\boldsymbol{\theta}_m = \frac{n_{k,v}^{\backslash m,n} + \boldsymbol{\eta}_v}{n_{k,\cdot}^{\backslash m,n} + \sum_{v'} \boldsymbol{\eta}_{v'}} \cdot \frac{n_{m,k}^{\backslash m,n} + \boldsymbol{\alpha}_k}{n_m^{\backslash m,n} + \sum_{k'} \boldsymbol{\alpha}_{k'}}, \tag{1.5}$$

where $n_{k,v}^{\backslash m,n}$ is the total number of counts allocated to topic k, excluding the nth gene v in sample m, and $n_{m,k}^{\backslash m,n}$ is the number of counts in sample m excluding the genes allocated to topic k, respectively. By utilizing these probability distributions, latent topics are sampled, and genes are assigned. The pseudo code for collapsed Gibbs sampling is described in Algorithm1.

The posterior distribution of LDA can be divided into two parts [19,23], and $p(\boldsymbol{Z} | \boldsymbol{\alpha})$ and $p(\boldsymbol{G} \mid \boldsymbol{Z}, \boldsymbol{\eta})$ are expressed as follows:

$$p(\mathbf{Z} | \boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\alpha}) \equiv \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K Z_k^{\alpha_k - 1} \tag{1.6}$$

$$p(\mathbf{G} | \mathbf{Z}, \boldsymbol{\eta}) = \frac{\Gamma\left(\sum_{k=1}^K n_k + \alpha_k\right)}{\prod_{k=1}^K \Gamma(n_k + \alpha_k)} \prod_{k=1}^K Z_k^{n_k + \alpha_k - 1}, \tag{1.7}$$

where $\Gamma(\cdot)$ is the gamma function, and n_k is the total number of genes allocated to topic k. Through Equations 1.6 and 1.7, the sum of the log-likelihood is calculated and the convergence with iteration is monitored.

The LDA algorithm provides the topic distribution for each sample $\boldsymbol{\theta}_m$ and the gene distribution

for each topic $\boldsymbol{\varphi}_k$ as output. In the context of cell-type deconvolution, the topics in the obtained $\boldsymbol{\theta}_m$ correspond to the cell types that make up the sample.

Marker gene name based guided LDA algorithm

The LDA algorithm is an unsupervised learning algorithm that poses a challenge in identifying the cell type corresponding to a particular topic. To address this issue, guided LDA models have been proposed, which incorporate prior information to guide the topics and specify a direction towards the generation process of the LDA model [24,25]. In the context of deconvolution, the marker gene name v of the cell corresponding to topic k' is used as prior information to guide the LDA algorithm as follows:

$$\varphi_{k,v} = \begin{cases} 1, & k = k' \\ 0, & k \neq k' \end{cases}, \tag{2.1}$$

where the general concept of this process is to initialize the topic-gene distribution and to link topics to cells and allow identification of specific cell proportions. During the learning process, the contribution of marker genes to the corresponding topic changes. Therefore, for guided marker genes, it is also possible to probabilistically avoid updating $\boldsymbol{Z}_{m,n}$ using a parameter.

Selection of genes for analysis

Consider a measured bulk gene expression matrix $\boldsymbol{Y} \in \boldsymbol{R}^{\boldsymbol{L}\times\boldsymbol{M}}$ for L types of genes across M samples. In our proposed algorithm, collapsed Gibbs sampling is repeated N times, the sum frequency of L types of gene expression in all M samples. To minimize computational costs, it is essential to carefully select genes for analysis. In addition to genes exhibiting immune-related behavior to guide the topics, we also selected genes with high coefficients of variation among samples for a study of interest. This was done to adapt the model to the study. Note that it is assumed that variation in the transcriptome may arise from changes in cell proportion or from simple expression changes, such as those caused by perturbation. However, to avoid confounding factors, we eliminated genes that were outliers greater than 2σ from the log-normal distribution of gene expression levels.

Additional topics for tissue data analysis

Tissues are composed of a diverse range of cell populations. In addition to the guided topics that allow for cell identification, we also observe the effects of other topics, such as unexpected cell types or confounding between experiments. To account for these factors, we assume that the K topics that make up the sample are the sum of the K_g guided topics of interest and the K_e empty additional topics with no prior information. To determine the optimal number of additional topics K_e, we have developed a recursive algorithm that tests gene contribution to each topic. After adding k_e empty topics and

performing deconvolution, we can determine the gene contribution to each topic φ_k for k = 1, 2, …, k_e. For each topic, we extract the genes whose contribution to the topic is in the top $1 / k_e$, and finally pool N_g genes. For the contribution profile of N_g genes across k_e samples, we obtained Pearson correlation matrix $\boldsymbol{P} \in \boldsymbol{R}^{k_e \times k_e}$ between the topics. When $\boldsymbol{P}$ is significantly negative for all elements, each of the k_e topics is considered independent. The value of k_e is increased until each topic can no longer be considered independent, and the final K_e is determined.

Ensemble strategy for robust estimation

The proportions of cell types in a sample are represented by the sample-topic distribution θ_m $(m = 1, \cdots, M)$. Due to the dependence of the Gibbs sampling process on random variables, a single trial may not be sufficient for accurate estimation. To address this, we employed the median ensemble strategy, which integrates θ_m from multiple random seeds and ensures that the cell type proportions sum to 1.

Functional analysis for the additional topics

We can obtain $K_e \times En$ topic-gene distributions for unknown topics by assembling the output of the models considering K_e additional topics En times. K-means clustering was employed to determine the contribution of L genes to each potential additional topic. Subsequently, enrichment analyses of the top contributing L/K_e genes with Gene Ontology (GO) [26] and KEGG pathway [27] were performed utilizing the Fisher's exact test. We also performed single-sample gene set enrichment analysis (ssGSEA) on the gene contribution values.

Data preparation

To assess the performance of GLDADec in estimating gene expression levels in blood and tissue samples, we selected datasets that included corresponding bulk transcriptome and immune cell proportions determined by flow cytometry. Additionally, we obtained human clinical data for practical real-world data application. GLDADec takes a gene expression matrix $\boldsymbol{Y} \in \boldsymbol{R}^{\boldsymbol{L} \times \boldsymbol{M}}$ for L genes across M samples as input, and utilizes non-log linear scale data derived from the bulk transcriptome. A summary of the datasets used is provided in Supplementary Table S1. See Supplementary Noste S1 for details on the experimental conditions, including preprocessing of each dataset.

Blood-derived benchmarking data

The PBMC microarray, PBMC/PMN RNA-Seq, and whole blood RNA-Seq datasets were obtained from the NCBI GEO database, with accessions GSE65133, GSE107572, and GSE60424 respectively. These blood-derived samples are commonly utilized to evaluate deconvolution methods [13,14,28–31]. A list of typical marker genes for the cells measured in the benchmark datasets was established using domain knowledge and is provided in Supplementary File S1.

Tissue data with perturbation

To assess the estimation of immune cell trafficking in tissues, we retrieved bulk RNA-Seq data from mice and rats with accessions GSE237801 and GSE239996. These data encompass the tissue transcriptome from drug-induced liver injury models and the principal immune cell proportions in the tissues. Marker genes for cell types linked to mouse liver were obtained from CellMarker [32]. Cleansing was performed by integrating homologous cell names and eliminating abstract cell names, and finally marker genes specific to the 34 cell types were defined (Supplementary File S2). As there is no established database for marker gene names in rats, we obtained classical cell-type markers that are highly expressed in rats, mice, and humans, as reported in the study by Natasha et al [33].

Clinical data

To assess the utility of GLDADec in the context of human clinical data, we obtained three distinct types of tumor datasets, comprising 1052 breast invasive carcinoma (BRCA) samples, 577 lung adenocarcinoma (LUAD) samples, and 408 liver hepatocellular carcinoma (LIHC) samples, from the GDC data portal [34]. Marker genes for the cell types associated with each human tissue were obtained from the CellMarker database [32].

Comparisons with other deconvolution methods

We conducted a comparison of the performance of GLDADec with seven competing deconvolution methods, including FARDEEP, EPIC, CIBERSORTx, DCQ, NNLS, RLR, and Elastic Net, utilizing their default settings where applicable [7–10]. Note that the input matrix for each method was normalized separately. Specifically, a log normalized gene expression matrix was used for DCQ and FARDEEP, while a non-log linear scaled matrix was utilized for the remaining methods. See Supplementary Noste S2 for further details.

Results

Benchmarking with human blood samples

To begin with, our algorithm was evaluated using three datasets comprising bulk transcriptome data from human blood (GSE65133, GSE107572, GSE60424)[13,28,29], along with the corresponding proportions of immune cell types determined through flow cytometry. The marker gene names specific to each cell type were defined using domain knowledge, and 100 genes with large coefficients of variation (CV) between samples were included for analysis (as outlined in Supplementary File S1). Our proposed method, which utilizes marker gene names based on domain knowledge as prior information, achieved highly accurate predictions (Pearson correlations ranging from 0.39 to 0.99), considering each cell type individually (Figure 2A, Figure S1). To ensure that these estimates were robust and independent of the number of added high CV genes and hyperparameters, a sensitivity analysis was conducted (see *Hyperparameter sensitivity analysis* section).

Additionally, we compared our proposed GLDADec method to existing reference-based methods, including FARDEEP, EPIC, CIBERSORT, DCQ, NNLS, RLR, and ElasticNet. The proposed method demonstrated a high Pearson correlation for each cell estimation across all benchmark datasets. Notably, for the dendritic cells of GSE107572, the estimation performance was significantly improved (Figures 2B-C, S2). The detailed experimental conditions for using existing reference-based methods are provided in Supplementary Note S2.

Impact of introducing additional topics

The transcriptome of tissues consists of heterogeneous cell populations, including cells that are not typically expected to be present in the tissue. To account for these unknown cell types, we performed modeling that treated them as additional topics. We validated the impact of introducing additional topics by evaluating immune cell trafficking in tissues with perturbation, using liver bulk RNA-Seq data from various drug-induced liver injury mouse models. We compared modeling with and without additional topics, considering four known cell types (neutrophils, monocytes, NK cells, and Kupffer cells).

The optimal number of additional topics increased with the number of genes analyzed for each acetaminophen (APAP) and alpha-naphthyl isothiocyanate (ANIT) treatment group (Figure 3A and S3A). Modeling with additional topics suppressed the variance of the estimates in the ensemble for several cell types, improving the performance of estimating the proportion of each cell type (Figure 3B and S4). In particular, the estimation of Kupffer cells with APAP administration was greatly improved (Figure 3C and S3B). To gain further biological insight into the additional topics, we conducted Gene Ontology (GO) [26] and KEGG pathway [27] analysis for the genes with large contribution to each additional topic. The

significant biological processes were mainly related to metabolism and biosynthesis, consistent with a major biological role for liver tissue and reflecting the influence of potential transcriptomes such as hepatocytes (Figure 3D, Table 1, and Table S2).

These results indicate that adding topics can improve the performance of guided cell estimation and allows the aggregation of biological functions that exist in the background of the target tissue as unknown topics.

Comprehensive cell type analysis for mouse data

In recent years, databases containing marker genes have become prevalent and easily accessible [32,35]. By obtaining a data-driven approach to obtain marker gene names for a diverse array of cell types, we can accurately estimate the proportions of a comprehensive cell type.

In this study, we curated marker genes in mouse liver obtained from CellMarker [32] and defined marker genes specific to liver-related cell types (Supplementary File S2). Using these marker genes and applying the proposed method to liver tissue during drug-induced liver injury, we estimated the proportion changes for a wide range of 26 cell types (Figure 4A and S5A). Notably, GLDADec allows the estimation of cell types such as hepatocytes and vascular smooth muscle cells (VSMCs), which have rarely been considered in conventional deconvolution methods. Pearson correlation was evaluated for four cell types validated by flow cytometry: neutrophils, monocytes, NK cells, and Kupffer cells. These cell types are widely recognized as playing a significant role in the pathophysiology of drug induced liver injury. As shown in Figures 4B and S5B, the proposed method accurately estimated the change in immune cell proportion with perturbation. In addition, although there were no ground-truths measured by flow cytometry for cell types such as hepatocytes and VSMCs, they showed significant positive or negative correlations with blood biochemistry values such as alanine aminotransferase (ALT) and aspartate aminotransferase (AST) (Tables S2, S3 and Figure S6). These results suggest that accurate estimates that reflect individual responses can be achieved for a wide range of cell types.

To assess the robustness of the proposed method to prior knowledge, we altered the marker genes used. Differentially expressed genes (DEGs) between LM13, 13 representative cell types of mouse liver, derived from transcriptome data of each cell type were deemed as marker genes and exhibited consistent superior estimation performance (Figure S7). Additionally, we evaluated the performance of existing reference-based methods utilizing DEGs, and the proposed method surpassed that of existing methods (illustrated in Figure 4C and S5C).

Application to rat data with poor marker information

Owing to the extensive use of rats in toxicology, there is plenty of toxicogenomic data for rats, a valuable resource for data analysis in drug and chemical safety assessments [18,36,37]. On the other hand, the cell marker information of rats was much lower than that of the mice. Our previous work showed that when using existing reference-based deconvolution, mouse-derived reference expressions are not extrapolatable, and rat-specific references should be used. However, databases containing gene expression profiles specific to each cell type in the rat are not as abundant as those in the mouse and must be obtained independently, which is costly and time-consuming.

In this section, we evaluate the performance of the proposed method using marker gene names defined in various scenarios. The method of defining the names of marker genes used to evaluate immune cell trafficking in the rat liver injury model was as follows:

1. Classical cell type markers for rats, mice, and humans reported by Natasha et al. [33]
2. Marker genes in the mouse liver obtained from CellMarker.
3. Names of DEGs derived from mouse LM6.

We found that the estimation performance was excellent when either was used as prior information (Figure 5A). Additionally, the estimation performance of the proposed method using each marker outperformed that of the existing reference-based method using mouse LM6, which is a set of representative immune cell types widely used in deconvolution (Figure 5B). These results suggest that the proposed method, which does not depend on the expression level of the reference gene, effectively achieves cell type proportion estimation using a priori information on the marker gene names that are conserved across species.

Furthermore, our analysis of the immune response in rat liver injury using the GLDADec database provides a more extensive insight into the immune response in rat liver injury, utilizing the mouse marker gene names. By obtaining the same marker genes examined in mouse liver injury, we were able to estimate the comprehensive cell type proportions in rat liver injury and compare the profile changes between mice and rats (Figures S8A and B). Our findings revealed a striking similarity in the changes of cell proportions in mice and rats, such as an increase in neutrophils and monocytes and a decrease in hepatocytes, following the administration of acetaminophen (Figures 5C and S8C). Additionally, we identified distinct immune cell trafficking patterns, including increased natural killer T (NKT) cells in mice but decreased in rats (Figure 5D). These conclusions indicate that our proposed method is a valuable tool for identifying similarities and differences in immune cell trafficking between species through comprehensive estimation of cell type proportions.

Application GLDADec to tumor samples

The tumor tissue is composed of various cell types, including infiltrating immune cells, stromal and vascular cells, and subclonal cancer cells, as well as other cell types [38]. To validate our prediction in tumors, we applied GLDADec to 2037 tumor samples from The Cancer Genome Atlas (TCGA) for three tumor types: breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), and liver hepatocellular carcinoma (LIHC) [39–42]. We used marker gene names obtained from CellMarker as prior information for each cancer subtype in the background tissues. Our proposed method stratified the cancer subtypes by comprehensively estimating the proportion of various cell types, including tissue-specific cells such as ciliated cells in LUAD samples and hepatocytes in LIHC samples (Figure 6A).

To assess the refined stratification performance more comprehensively, we examined the differences in estimated cell proportions for each breast cancer subtype, as reported by Brian et al. [43] (Figure 6B). As anticipated, the proportion of luminal cells was found to be higher in subtypes LumA and LumB (Figure 6C). Moreover, we observed a distinct pattern of immune cell infiltration, particularly monocytes, B cells, and plasma cells, in the Her2 subtype (Figure 6D and S9). This suggests that subtype classification based on cell proportions estimated by GLDADec is feasible and may offer novel insights into the unique migration patterns of immune cells within each subtype.

Next, we examined the relationship between the predicted proportion of cell types and patient survival. Tumor samples were categorized according to the estimated proportion of specific cell types, and Cox proportional hazards regression was employed to determine survival rates between the top and bottom 25% of samples. Our analyses revealed that patients with lung adenocarcinoma (LUAD) exhibiting a high level of B cell infiltration had worse overall survival (P = 0.012), while those with liver hepatocellular carcinoma (LIHC) demonstrated improved survival (P = 8.13e-03) (Figure 6E). These findings indicate that B cells contribute differently to the prognosis and treatment of breast and liver cancers, which is consistent with previous reports [44,45]. Moreover, our study found that the infiltration of immune cells, such as γδT and CD4+ T cells, was associated with better clinical outcomes in breast cancer, which supports previous studies [46,47] (Figure S10). In addition, GLDADec can estimate the proportion of various cell types with marker gene names, and it is expected to provide novel insights. For instance, we reported that the loss of alveolar type II cells may be the cause of the poor prognosis of LUAD (Figure 6F).

Hyperparameter sensitivity analysis

The benchmark dataset was utilized to carry out sensitivity analyses for the hyperparameters $\boldsymbol{\alpha}$ for sample-topic distribution, $\boldsymbol{\eta}$ for topic-gene distribution, as well as the number of high CV genes and the

number of ensemble attempts. It was observed that GLDADec exhibits robustness with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ under conditions between 0 and 1, which are commonly used in LDA [48,49]. Under conditions that allow for more than one, the estimates were found to be less dependent on $\boldsymbol{\alpha}$ and more susceptible to changes in $\boldsymbol{\eta}$ (Figure S11A). In terms of the number of high CV genes to add, it was observed that gamma-delta T cells (γδT) exhibited better estimation performance when a higher number of genes were added; however, this effect was limited for many other cell types (Figure S11B). Regarding the number of trials to ensemble, there was little effect on the human blood-derived benchmark dataset, while more accurate estimation performance was obtained by increasing the number of trials in the analysis on perturbed mouse and rat tissues (Figure S11C and D).

Discussion

We introduce GLDADec, a marker gene name-based deconvolution approach, to estimate the cell type proportions present in a heterogeneous sample. In the process of inferring the topic distribution of each sample, we utilize marker gene names specific to each cell type as partial prior information to guide the cell name to the topic. Additionally, we can detect and incorporate unknown potential topics that characterize the sample, alongside the guided topics, simultaneously. The estimated topic distribution can be utilized to calculate the proportion of cell types from the bulk RNA-Seq dataset.

A notable advantage of GLDADec is the ease of accessing the name of the marker gene used as prior information . Although gene expression profiles of immune cells can typically be obtained in humans and mice, this is not always feasible for species such as rats that are used in limited areas or for tissue-specific parenchymal cells [37]. As a result, reference-based methods are dependent on the amount of prior information available for the scope of their analysis. However, marker gene names are uniformly accessible from databases, and there are also classical markers that are conserved across species, enabling a far more extensive range of prior information to be utilized in terms of the number of cell types and species to be analyzed.

One of the benefits of this method is its ability to infer additional topics other than guided topics, which may be useful in accounting for the impact of unknown cell types or removing confounding factors such as batch effects. In our analysis of liver samples from a drug-induced liver injury mouse model, we observed that the estimation of Kupffer cells was significantly improved by considering additional topics, whereas the benefits for neutrophils were minimal (Figure 3C). This may be due to the variance in estimated values during the ensemble for each cell type. For neutrophils, the estimated values were robust

even without considering additional topics, suggesting that the variance in neutrophils was small and thus the impact of additional topics on the cell type was limited (Figure S5A). Additionally, we found that the additional topics not only reduced the variance of guided cell type proportions but also reflected biological characteristics, such as enrichment in metabolic and biosynthetic pathways (Figure 3D), which is consistent with the biological role of the liver. These topics likely reflect the influence of the transcriptome derived from hepatocytes, which constitute the majority of liver tissue.

Applying GLDADec, we quantified the fraction of diverse cell types within tumor samples registered in the TCGA database. We focused on minor cell types that were not typically identified through conventional approaches. For instance, we revealed the significant association between the presence of alveolar type II cells and unfavorable clinical outcomes for patients with LUAD (Figure 6F). These findings possess the potential to facilitate subsequent analyses and contribute novel perspectives.

It should be noted that the cell types covered by GLDADec are contingent upon the extent of knowledge of cell types and their marker genes that have been accumulated in the field of life sciences. In this study, we obtained marker gene names from a database, CellMarker, which is well-characterized, yet does not register eosinophils, a cell type that plays a protective role in acute liver injury and is liver-related. Therefore, for a more comprehensive analysis, it may be necessary to integrate multiple databases that store marker genes to consolidate information. Furthermore, considering recent advances in large language models, gathering marker gene information from scientific papers could be helpful to extract the marker gene names of cell types resident in a tissue of interest. While social acceptance is crucial as knowledge, such knowledge miners could prove to be powerful partners of the proposed method.

Key Points

- We introduce GLDADec, a novel bulk deconvolution method leveraging marker gene names as partial prior information for estimating cell type proportions.
- GLDADec adopts a semi-supervised learning algorithm, utilizing cell type marker genes to address challenges present in both conventional reference-based and reference-free methods simultaneously.
- GLDADec demonstrates strong estimation performance and biological interpretability as evidenced by benchmarking across blood-derived and perturbed tissue datasets.

- We utilized GLDADec on TCGA tumor samples, conducting cancer subtype stratification and survival analysis, showcasing its utility in clinical data analysis.

Declarations

Ethics approval and consent to participate

The studies reported in this article were performed in accordance with the guidelines provided by the Institutional Animal Care Committee (Graduate School of Pharmaceutical Sciences, the University of Tokyo, Tokyo, Japan).

Consent for publication

Not applicable.

Availability of data and materials

Code, models, and data are available at https://github.com/mizuno-group/GLDADec.

Competing interests

The authors declare that they have no conflicts of interest.

Funding

This work was supported by the JSPS KAKENHI Grant-in-Aid for Scientific Research (C) (grant number 21K06663) from the Japan Society for the Promotion of Science, Takeda Science Foundation, and Mochida Memorial Foundation for Medical and Pharmaceutical Research.

Authors' information

Affiliations

Graduate School of Pharmaceutical Sciences, the University of Tokyo, Bunkyo-ku, Tokyo, 113-0033, Japan

Iori Azuma, Hiroyuki Kusuhara

Graduate School of Pharmaceutical Sciences, the University of Tokyo, Bunkyo-ku, Tokyo, 113-0033, Japan, tadahaya@gmail.com

Tadahaya Mizuno

Contributions

IA: Conceptualization, Data curation, Formal analysis, Methodology, Software, Investigation, Writing – Original draft, Visualization.

TM: Conceptualization, Resources, Supervision, Project administration, Writing – Original draft, Writing – Review and editing, Funding acquisition.

HK: Writing – Review and editing.

References

1. Lee WM. Drug-Induced Hepatotoxicity. New England Journal of Medicine 2003; 349:474–485

2. Egeblad M, Nakasone ES, Werb Z. Tumors as Organs: Complex Tissues that Interface with the Entire Organism. Dev Cell 2010; 18:884–901

3. Clough E, Barrett T. The Gene Expression Omnibus Database. 2016; 93–110

4. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013; 45:580–585

5. Lachmann A, Torre D, Keenan AB, et al. Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun 2018; 9:1366

6. Wu AR, Neff NF, Kalisky T, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods 2014; 11:41–46

7. Hao Y, Yan M, Heath BR, et al. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. PLoS Comput Biol 2019; 15:e1006976

8. Racle J, de Jonge K, Baumgaertner P, et al. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife 2017; 6:

9. Steen CB, Liu CL, Alizadeh AA, et al. Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx. Methods Mol Biol 2020; 2117:135–157

10. Altboum Z, Steuerman Y, David E, et al. Digital cell quantification identifies global immune cell dynamics during influenza infection. Mol Syst Biol 2014; 10:

11. Qin Y, Zhang W, Sun X, et al. Deconvolution of heterogeneous tumor samples using partial reference signals. PLoS Comput Biol 2020; 16:e1008452

12. Kang K, Meng Q, Shats I, et al. CDSeq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. PLoS Comput Biol 2019; 15:e1007510

13. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015; 12:453–457

14. Racle J, de Jonge K, Baumgaertner P, et al. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife 2017; 6:

15. Chu T, Wang Z, Pe'er D, et al. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. Nat Cancer 2022; 3:505–517

16. Chen Z, Wu A. Progress and challenge for computational quantification of tissue immune cells. Brief Bioinform 2021; 22:

17. Titus AJ, Gallimore RM, Salas LA, et al. Cell-type deconvolution from DNA methylation: a review of recent applications. Hum Mol Genet 2017; 26:R216–R224

18. Azuma I, Mizuno T, Morita K, et al. Investigation of the usefulness of liver-specific deconvolution method by establishing a liver benchmark dataset. NAR Genom Bioinform 2024; 6:

19. David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent dirichlet allocation. The Journal of Machine Learning Research 2003; 3:993–1022

20. Chauhan U, Shah A. Topic Modeling Using Latent Dirichlet allocation. ACM Comput Surv 2022; 54:1–35

21. Swapna LS, Huang M, Li Y. GTM-decon: guided-topic modeling of single-cell transcriptomes enables sub-cell-type and disease-subtype deconvolution of bulk transcriptomes. Genome Biol 2023; 24:190

22. Griffiths TL, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences 2004; 101:5228–5235

23. Buntine W. Estimating Likelihoods for Topic Models. 2009; 51–64

24. Watanabe K, Baturo A. Seeded Sequential LDA: A Semi-Supervised Algorithm for Topic-Specific Analysis of Sentences. Soc Sci Comput Rev 2023;

25. Zhou S, Kan P, Huang Q, et al. A guided latent Dirichlet allocation approach to investigate real-time latent topics of Twitter data during Hurricane Laura. J Inf Sci 2023; 49:465–479

26. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000; 25:25–29

27. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000; 28:27–30

28. Finotello F, Mayer C, Plattner C, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med 2019; 11:34

29. Linsley PS, Speake C, Whalen E, et al. Copy Number Loss of the Interferon Gene Cluster in Melanomas Is Linked to Reduced T Cell Infiltrate and Poor Patient Prognosis. PLoS One 2014; 9:e109760

30. Menden K, Marouf M, Oller S, et al. Deep learning–based cell composition analysis from tissue expression profiles. Sci Adv 2020; 6:

31. Andrade Barbosa B, van Asten SD, Oh JW, et al. Bayesian log-normal deconvolution for enhanced in silico microdissection of bulk gene expression data. Nat Commun 2021; 12:6106

32. Hu C, Li T, Xu Y, et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids Res 2023; 51:D870–D876

33. Qaisar N, Arowosegbe A, Derr AG, et al. Type I IFN–Driven Immune Cell Dysregulation in Rat Autoimmune Diabetes. Immunohorizons 2021; 5:855–869

34. Grossman RL, Heath AP, Ferretti V, et al. Toward a Shared Vision for Cancer Genomic Data. New England Journal of Medicine 2016; 375:1109–1112

35. Franzén O, Gan L-M, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database 2019; 2019:

36. Igarashi Y, Nakatsu N, Yamashita T, et al. Open TG-GATEs: a large-scale toxicogenomics database. Nucleic Acids Res 2015; 43:D921–D927

37. Morita K, Mizuno T, Azuma I, et al. Rat deconvolution as knowledge miner for immune cell trafficking from toxicogenomics databases. Toxicological Sciences 2023;

38. de Visser KE, Joyce JA. The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth. Cancer Cell 2023; 41:374–403

39. Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013; 45:1113–1120

40. Berger AC, Korkut A, Kanchi RS, et al. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. Cancer Cell 2018; 33:690-705.e9

41. Campbell JD, Alexandrov A, Kim J, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat Genet 2016; 48:607–616

42. Ally A, Balasundaram M, Carlsen R, et al. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell 2017; 169:1327-1341.e23

43. Lehmann BD, Jovanović B, Chen X, et al. Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. PLoS One 2016; 11:e0157368

44. Zhang Y, Yin X, Wang Q, et al. A novel gene expression signature-based on B-cell proportion to predict prognosis of patients with lung adenocarcinoma. BMC Cancer 2021; 21:1098

45. Zhang X, Zhang J, Gao F, et al. KPNA2-Associated Immune Analyses Highlight the Dysregulation and Prognostic Effects of GRB2, NRAS, and Their RNA-Binding Proteins in Hepatocellular Carcinoma. Front Genet 2020; 11:

46. Tosolini M, Pont F, Poupot M, et al. Assessment of tumor-infiltrating TCRV γ 9V δ 2 γδ lymphocyte abundance by deconvolution of human cancers microarrays. Oncoimmunology 2017; 6:e1284723

47. Peng Y, Yu H, Jin Y, et al. Construction and Validation of an Immune Infiltration-Related Gene Signature for the Prediction of Prognosis and Therapeutic Response in Breast Cancer. Front Immunol 2021; 12:

48. Griffiths TL, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences 2004; 101:5228–5235

49. Panichella A. A Systematic Comparison of search-Based approaches for LDA hyperparameter tuning. Inf Softw Technol 2021; 130:106411

Figures

Figure 1

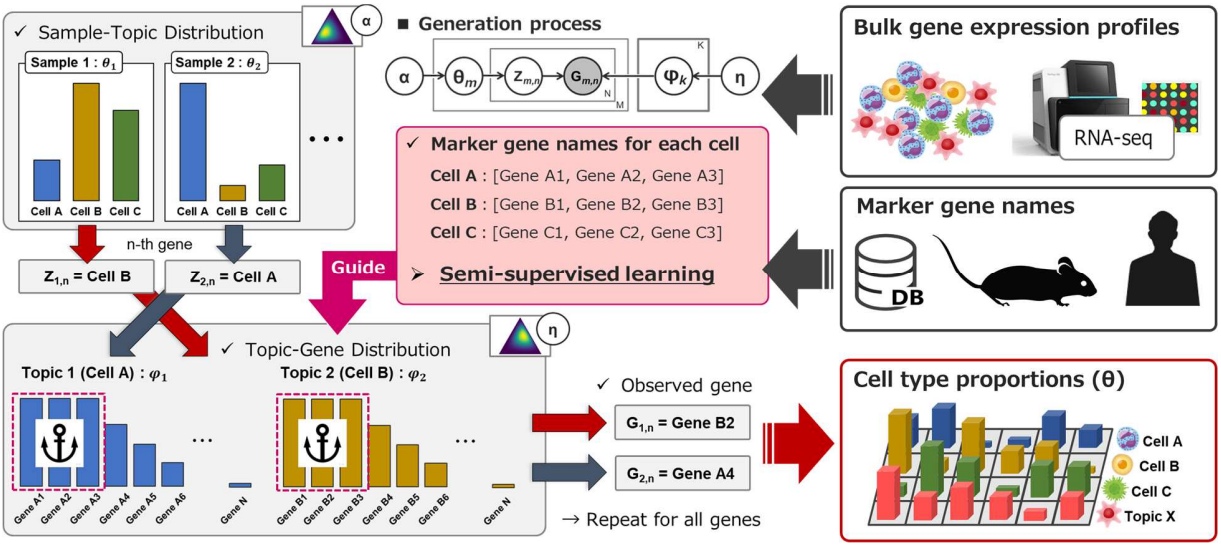


Fig. 1 Overview of GLDADec. The observed gene expression profiles are considered as bag-of-words. We extend the standard LDA generation process to incorporate semi-supervised learning, where the gene names specific to each cell (topic) serve as partial prior information to guide the process. By running GLDADec, we can obtain θ, which reflects the cell type proportions in each sample.

Figure 2

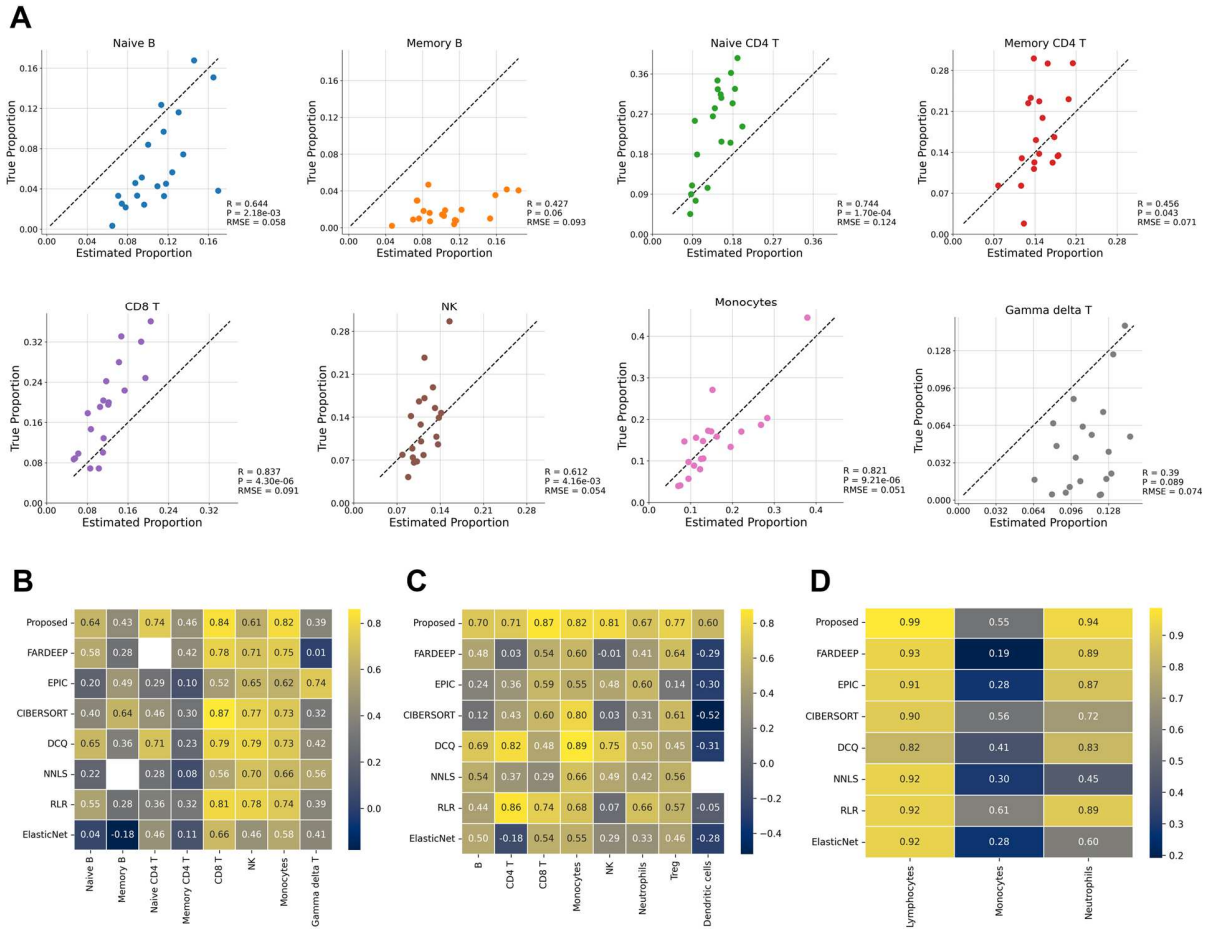


Fig. 2 Benchmarking with human blood samples. **(A)** Scatterplots comparing estimated proportions and measured immune cell proportions in GSE65133. **(B-D)** Pearson correlation values between the output proportions derived from various bulk deconvolution methods and the measured values in GSE65133, GSE107572, and GSE60424, respectively.

Figure 3

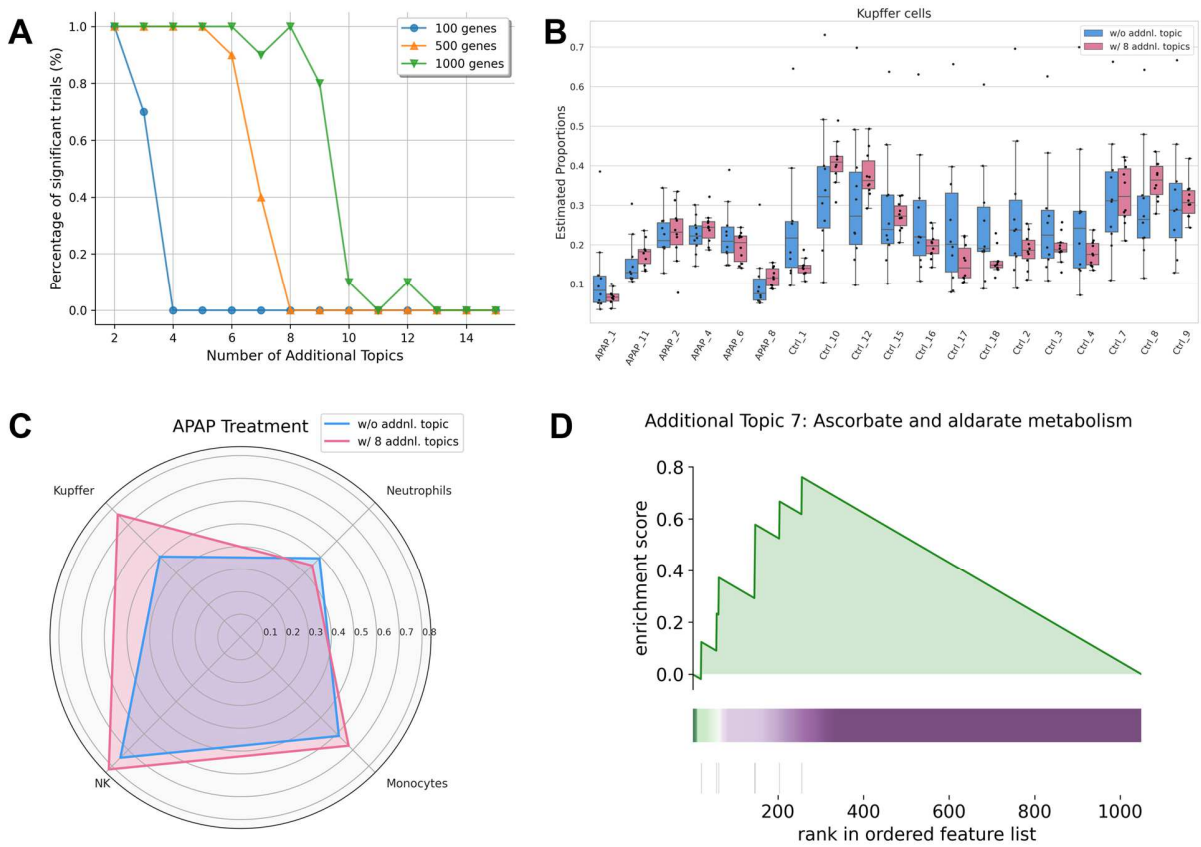


Fig. 3 Assessment of the value of incorporating additional topics in the analysis of tissue data affected by acetaminophen (APAP). **(A)** The relationship between the number of additional topics independent of each other and the number of genes to be analyzed. **(B)** Variance in estimates among ensemble members for each sample, both with and without taking additional topics into account. **(C)** Radar chart comparing the difference in estimated performance with and without the additional topics. The axis values signify the Pearson correlation between estimated and measured proportion of immune cells. **(D)** Gene set enrichment analysis (GSEA) was carried out on the ranked list of genes ranked according to their contribution to Topic 7. The colored band signifies the extent of each gene's contribution to the additional topic 7, with green representing a high contribution and purple indicating a low contribution. The bottom vertical black lines indicate the position of the genes that overlap with the pathway.

Figure 4

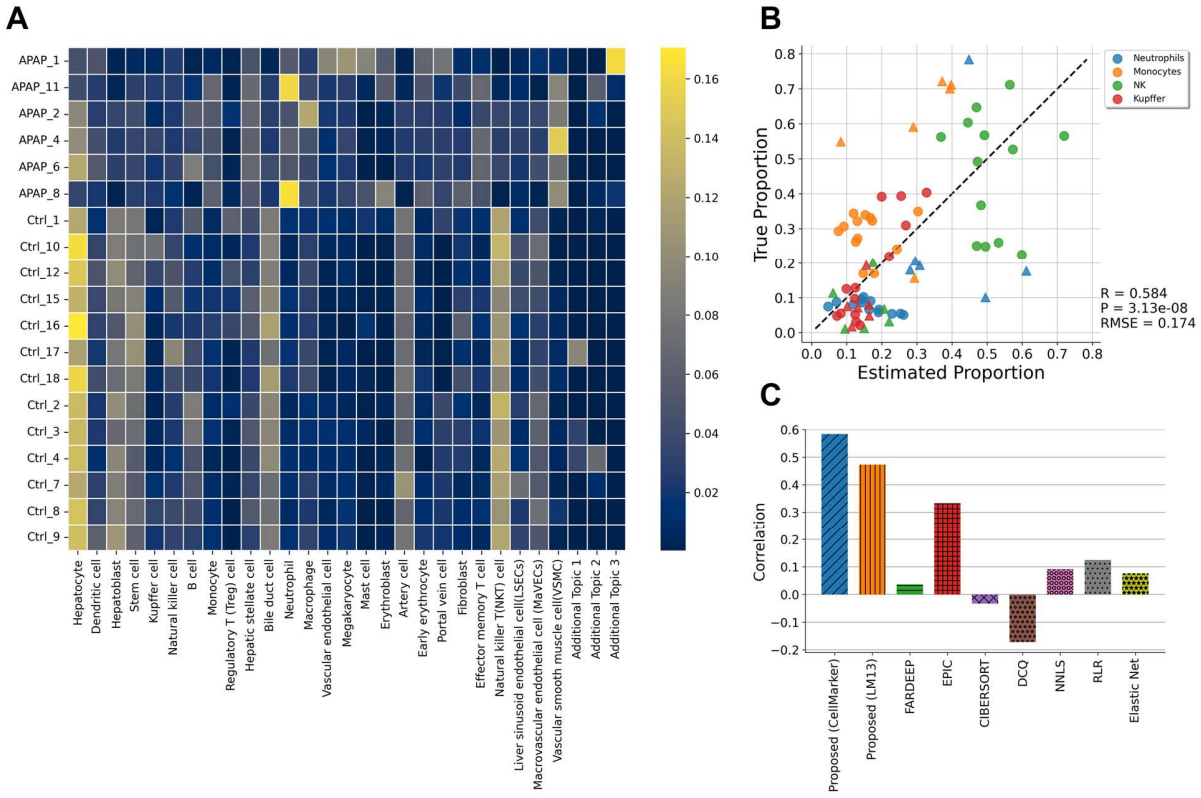


Fig. 4 Prediction of immune cell trafficking in mouse liver tissue perturbed by acetaminophen (APAP). **(A)** Comprehensive cell type proportions estimation using marker gene names collected in data-driven manner. **(B)** Scatterplot showing the estimated proportions of immune cells vs. measured values in the same dataset. The circles indicate the control samples, and the triangles indicate the samples treated with APAP. **(C)** Comparison of performance using Pearson correlation with other notable deconvolution techniques.

Figure 5

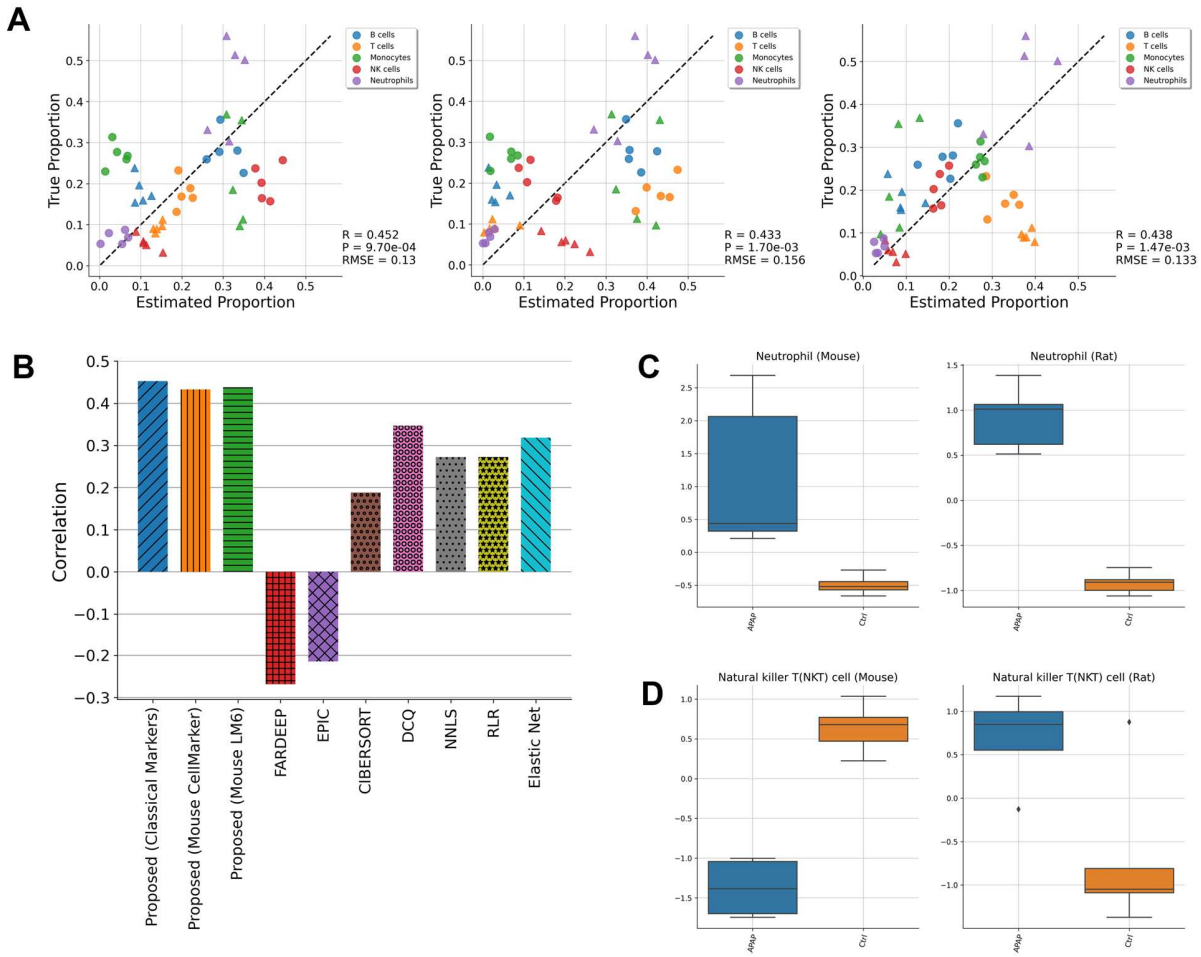
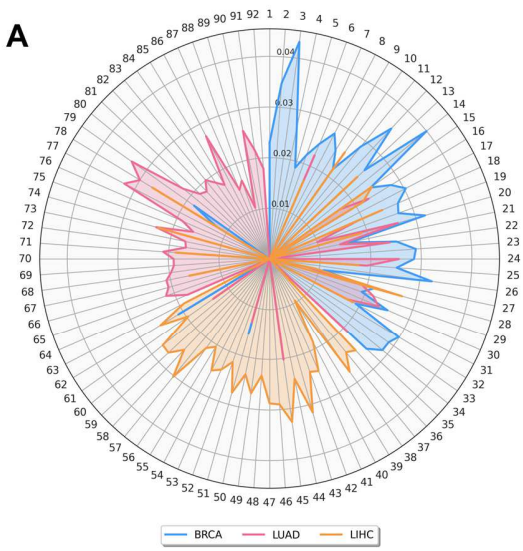


Fig. 5 Estimation on rat samples with poor marker information. **(A)** Scatterplots showing the estimated proportions of immune cells alongside the corresponding measured values within the same dataset. These graphics represent the outcomes of defining marker genes in three distinct scenarios: classical marker genes, mouse marker genes obtained from CellMarker, and differentially expressed genes derived from cell type specific gene expression profiles in mice. The circular symbols denote control samples, while the triangular symbols represent samples that have been treated with acetaminophen (APAP). **(B)** Comparison of performance using Pearson correlation with other notable deconvolution techniques. **(C-D)** Common and different patterns of immune cell trafficking in the liver between mice and rats treated with APAP.

Figure 6



1	Adipocyte	32	Basal epithelial cell	63	Vascular endothelial cell
2	GrB+ Regulatory B cell	33	Basal cell	64	Aerocyte
3	**Luminal cell**	34	Conventional T(Tconv) cell	65	Tissue-resident macrophage
4	Luminal epithelial cell	35	Eosinophil	66	Goblet cell
5	Lymphatic endothelial cell	36	Sinusoidal endothelial cell	67	Effector T(Teff) cell
6	Mammary stem cell	37	Resting regulatory T cell	68	Tumor-infiltrating regulatory T cell
7	Mast cell	**38**	**Portal vein epithelial cell**	69	Squamous cell
8	Melanocyte	39	Portal fibroblast	70	Smooth muscle cell
9	Mesenchymal stem cell	40	Memory T cell	71	Tumor-associated macrophage (TAM)
10	Monocyte	41	T follicular helper(Tfh) cell	72	Vascular cell
11	Muscle cell	42	T helper 1(Th1) cell	73	Alveolar cell Type 1
12	**Myoepithelial cell**	**43**	**Beta cell(β cell)**	74	Megakaryocyte
13	Myofibroblast	44	Naive B cell	75	Idiopathic pulmonary fibrosis cell
14	**Naive T(Th0) cell**	45	T helper 17(Th17) cell	76	Clara cell
15	Natural killer cell	46	T helper cell	**77**	**Ciliated cell**
16	Neutrophil	47	Epsilon cell	78	Macrophage
17	Paget cell	48	Cholangiocyte	79	Mesenchymal cell
18	Pericyte	49	Alpha cell (α cell)	80	Mesothelioma cell
19	Perivascular cell	50	Erythroid cell	81	Capillary cell
20	Plasma cell	51	CD8+ T cell	82	Lymphangioleiomyomatosis cell(LAMcore)
21	Red blood cell (erythrocyte)	52	Effector memory T cell	83	Alveolar cell Type 2
22	Regulatory T(Treg) cell	53	Hepatic stellate cell	84	Lobe cell
23	Suprabasal epithelial cell	54	Hematopoietic cell	85	Neuroendocrine cell
24	T helper(Th) cell	55	Kupffer cell	**86**	**Basophil**
25	Gamma delta(γδ) T cell	56	Liver stem cell	87	Pan-endothelial cell
26	Exhausted T(Tex) cell	**57**	**Hepatocyte**	88	Panvascular cell
27	others	58	Gallbladder cell	89	Ionocyte cell
28	B cell	59	Delta cell	**90**	**B-2 cell**
29	Cancer stem cell	60	Hepatoblast	91	Bronchial vessel cell
30	CD4+ T cell	61	Fibroblast	92	Lung cell
31	Dendritic cell	62	Cytotoxic T cell		

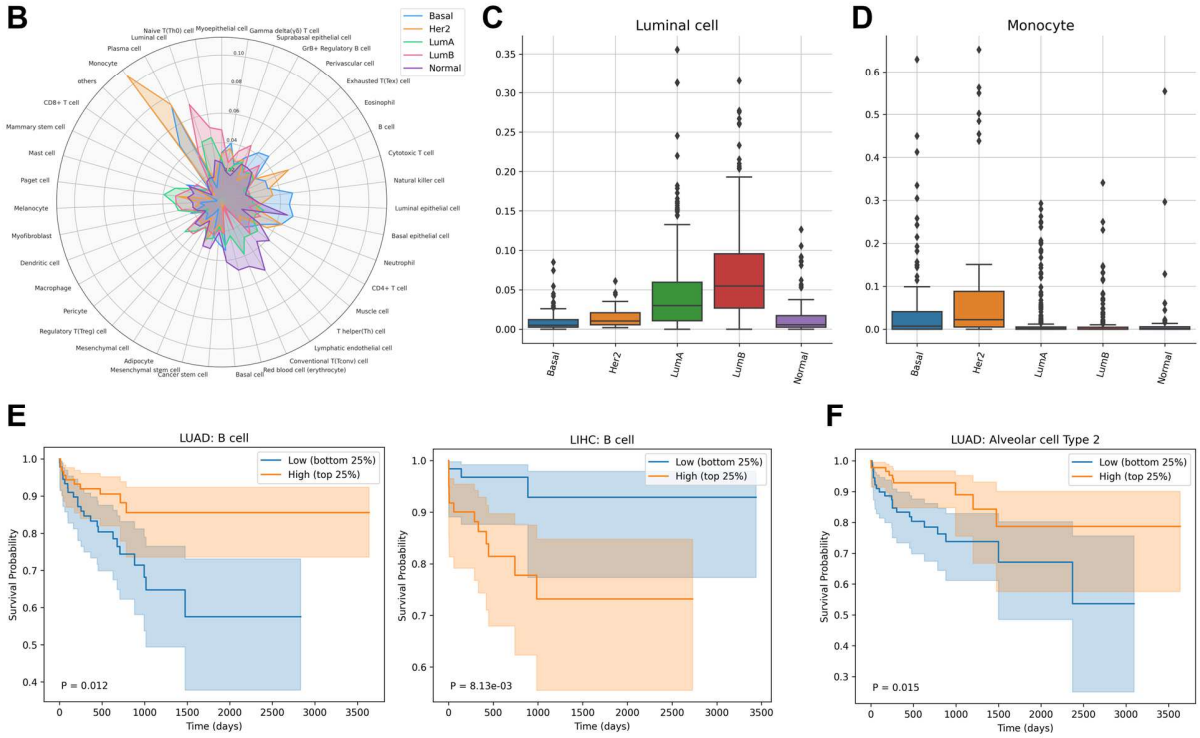


Fig. 6 Application of GLDADec to tumor samples. **(A)** Radar chart showing the estimated comprehensive cell type proportions in each tumor sample derived from BRCA, LUAD, and LIHC carcinomas. The table below the chart summarizes the number of cells for each cell type and corresponding vertex. The top three cell types, which are specific to each carcinoma and account for a substantial proportion, have been

emphasized. **(B)** Relative proportions of comprehensive cell types in different subtypes of BRCA. **(C-D)** Comparison of cells showing specific accumulation and infiltration between each subtype of BRCA. **(E-F)** Kaplan-Meier plots displaying the relationship between survival and the infiltration of B cells in LUAD and LIHC, as well as Alveolar cell type 2 in LUAD. The patients were divided into two groups based on the top and bottom 25th percentiles of the target cells, and the results were analyzed using the log-rank test. The transparent colors in the plots represent 95% confidence bands, and the *P*-values were computed using the log-rank test.

Tables

Table 1. KEGG pathway enrichment analysis. Significantly enriched pathway names are shown with Bonferroni adjusted *P*-values.

Topics	KEGG Pathway	Adjusted *P*-value	Overlap
2	Propanoate metabolism	0.005361	{'ACAT2', 'ALDH6A1', 'ABAT', 'PCCA', 'HIBCH', 'ACACA', 'ACACB'}
	Valine, leucine and isoleucine degradation	0.01992	{'ACAT2', 'ALDH6A1', 'ACADSB', 'ABAT', 'PCCA', 'HIBCH', 'ALDH3A2', 'ALDH1B1'}
	beta-Alanine metabolism	0.01992	{'ALDH6A1', 'ABAT', 'HIBCH', 'ALDH3A2', 'DPYD', 'ALDH1B1'}
7	Ascorbate and aldarate metabolism	0.01795	{'UGT1A6B', 'UGT2B35', 'UGT2A3', 'UGT2B34', 'UGT1A6A', 'ALDH1B1'}
	Pentose and glucuronate interconversions	0.02582	{'UGT1A6B', 'UGP2', 'UGT2B35', 'UGT2A3', 'UGT2B34', 'UGT1A6A'}
	Tryptophan metabolism	0.02582	{'MAOB', 'DDC', 'DHTKD1', 'IDO2', 'KYNU', 'AADAT', 'ALDH1B1'}
8	Tryptophan metabolism	0.01391	{'MAOB', 'DDC', 'DHTKD1', 'IDO2', 'KYNU', 'AADAT', 'ALDH3A2', 'ALDH1B1'}
	Ascorbate and aldarate metabolism	0.01391	{'UGT1A6B', 'UGT2B35', 'UGT2A3', 'ALDH3A2', 'UGT1A6A', 'ALDH1B1'}
	Chemical carcinogenesis	0.01391	{'CYP2B9', 'UGT1A6B', 'GSTA4', 'GSTM7', 'GSTM6', 'UGT2B35', 'CYP2B10', 'UGT2A3', 'GSTT3', 'UGT1A6A', 'GSTM4'}
	Metabolism of xenobiotics by cytochrome P450	0.01391	{'UGT1A6B', 'GSTM7', 'GSTA4', 'GSTM6', 'UGT2B35', 'UGT2A3', 'GSTT3', 'UGT1A6A', 'GSTM4'}
	Drug metabolism	0.01436	{'TK1', 'UGT1A6B', 'MAOB', 'GSTM7', 'GSTA4', 'GSTM6', 'UGT2B35', 'UGT2A3', 'DPYD', 'GSTT3', 'UGT1A6A', 'GSTM4'}

Algorithms

Algorithm 1.

Algorithm 1 : Overall Inference Process of GLDADec

```
Initialize α and η.
Stage1: initialize Z .
for m = 1, 2, ···, M do
    for n = 1, 2, ···, N_m do
        v ← G_{m,n}
        if v is contained in the markers of k'_{th} topic then
            Assign φ_{k',v} = 1.
        else
            Assign φ_{k',v} according to Dir(η).
        end if
    end for
end for

Stage2: Collapsed Gibbs sampling.
Set S: the number of samples.
for s = 1, 2, ···, S do
    for m = 1, 2, ···, M do
        for n = 1, 2, ···, N_m do
            Sample Z^{(s)}_{m,n} from (2.5).
            Update n^{(s)}_{k,v} and n^{(s)}_{m,k}.
        end for
    end for
end for
```

Supplementary figures

GLDADec: marker-gene guided LDA modelling for bulk gene expression deconvolution

Iori Azuma[*, 1], *Tadahaya Mizuno*[*, †, 2], *and Hiroyuki Kusuhara*[1]

[1]Graduate School of Pharmaceutical Sciences, the University of Tokyo, Bunkyo, Tokyo, 113-0033, Japan

[2]Graduate School of Pharmaceutical Sciences, the University of Tokyo, Bunkyo, Tokyo, 113-0033, Japan,

Electronic address: tadahaya@gmail.com

---

†Corresponding author: Tadahaya Mizuno, tadahaya@gmail.com

Graduate School of Pharmaceutical Sciences, the University of Tokyo, Bunkyo-ku, Tokyo, 113-0033, Japan

Tel: +81-3-5841-4771

*These authors contributed equally.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

FUNDING

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (C) (grant number 21K06663) from the Japan Society for the Promotion of Science, Takeda Science Foundation, and Mochida Memorial Foundation for Medical and Pharmaceutical Research.

Figure S1

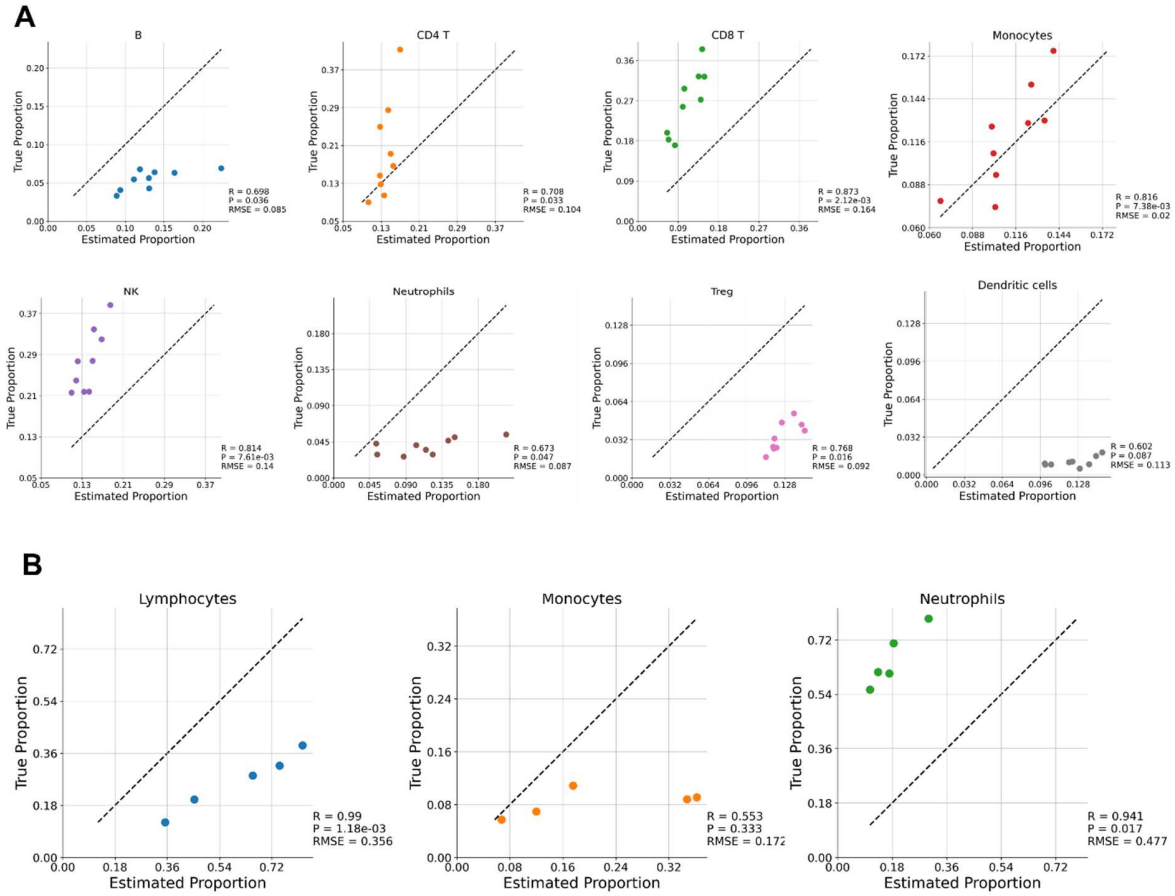


Figure S1. Benchmarking with human blood samples. Scatterplots comparing estimated proportions and measured immune cell proportions in **(A)** GSE107572 and **(B)** GSE60424.

Figure S2

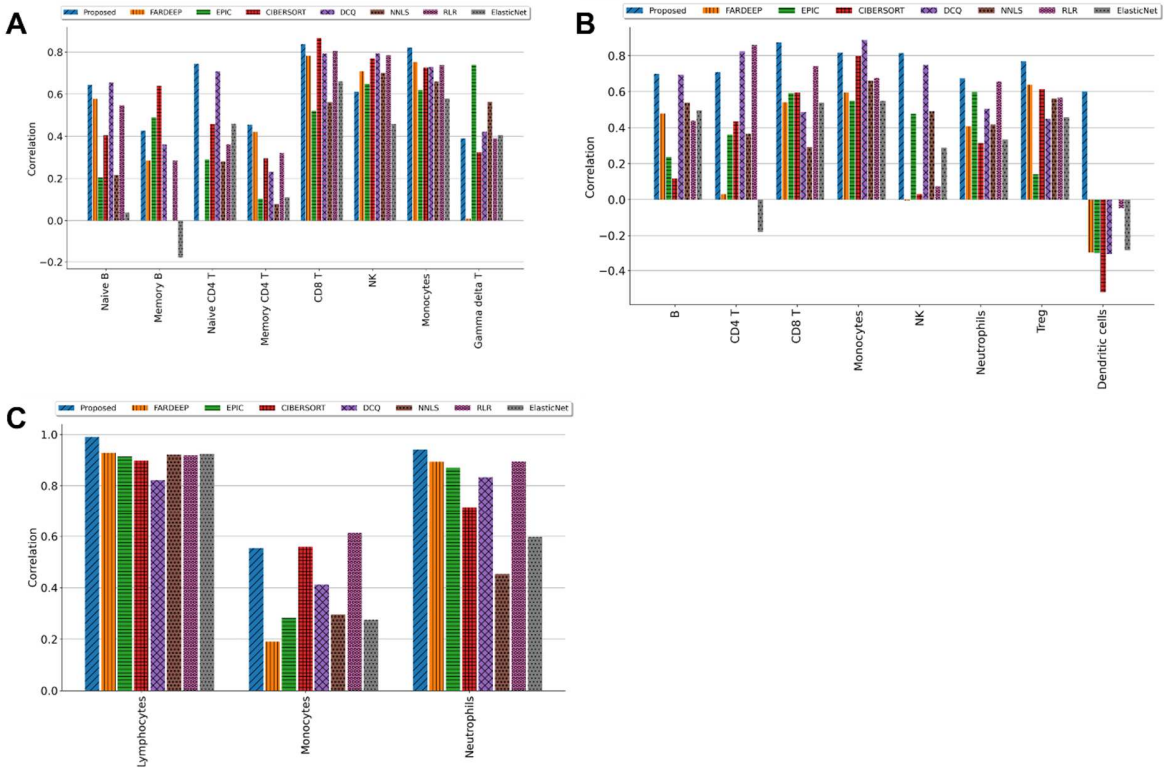


Figure S2. Pearson correlation values between the output proportions from the different bulk deconvolution methods and measured values in **(A)** GSE65133, **(B)** GSE107572, and **(C)** GSE60424, respectively.

Figure S3

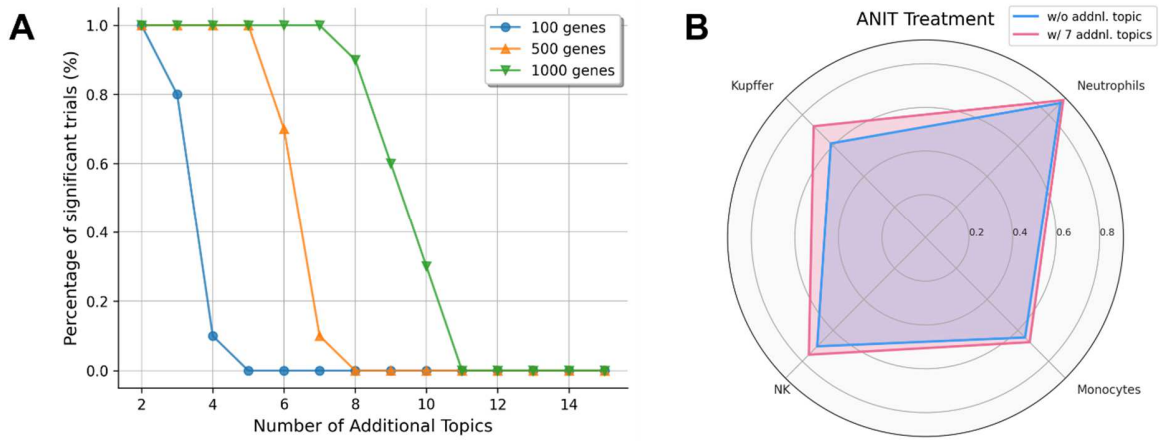


Figure S3. Evaluation of the usefulness of additional topics in alpha-naphthyl isothiocyanate (ANIT) perturbed tissue data. **(A)** The relationship between the number of additional topics that are significantly independent of each other and the number of genes to be analyzed. **(B)** Radar chart comparing the difference in estimated performance with and without the additional topics. The axis values indicate the Pearson correlation between estimated and measured proportion of each immune cells.

Figure S4

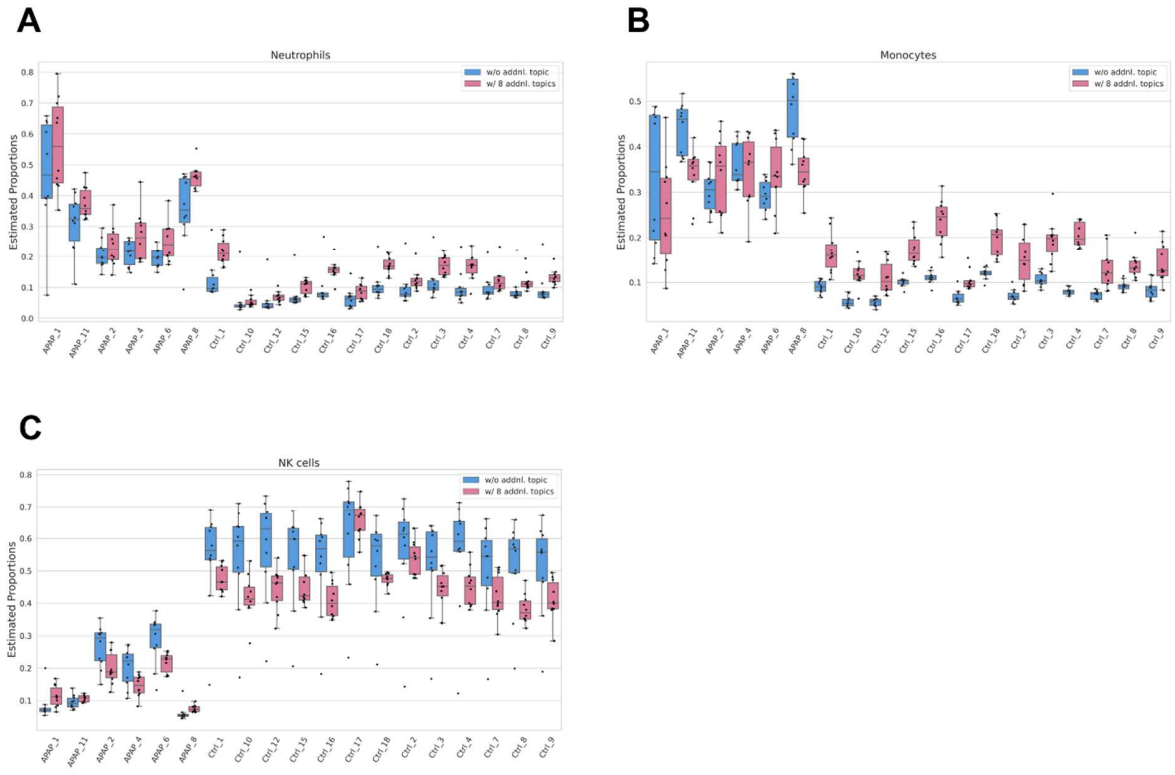


Figure S4. Variance of the estimate at ensemble in each sample with and without additional topics considered. Bar plots showing the estimated values of **(A)** neutrophils, **(B)** monocytes, and **(C**) Natural killer cells in each sample after APAP administration.

Figure S5

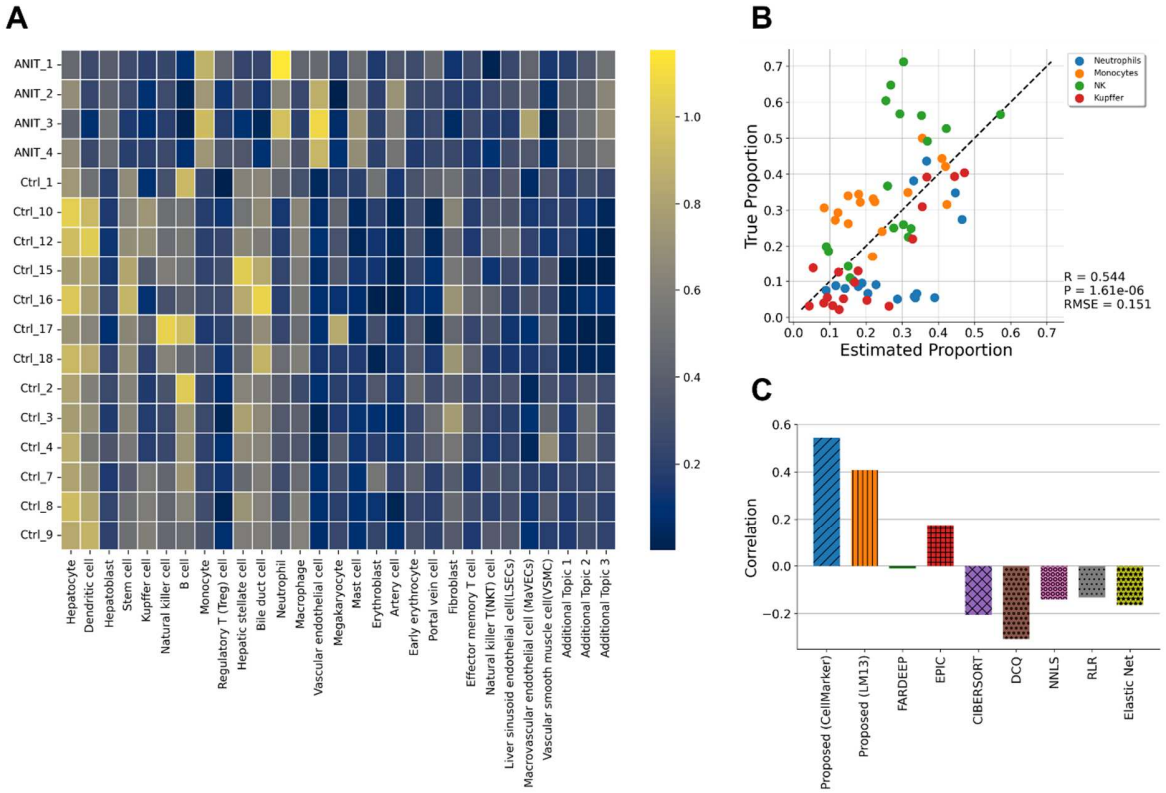


Figure S5. Prediction of immune cell trafficking in mouse liver tissue perturbed by alpha-naphthyl isothiocyanate (ANIT). **(A)** Comprehensive cell type proportions estimation using marker gene names collected in data-driven manner. **(B)** Scatterplot showing the estimated proportions of immune cells vs. measured values in the same dataset. Circles indicate control samples and triangles indicate samples treated with ANIT. **(C)** Performance comparison based on Pearson correlation with other representative deconvolution methods.

Figure S6

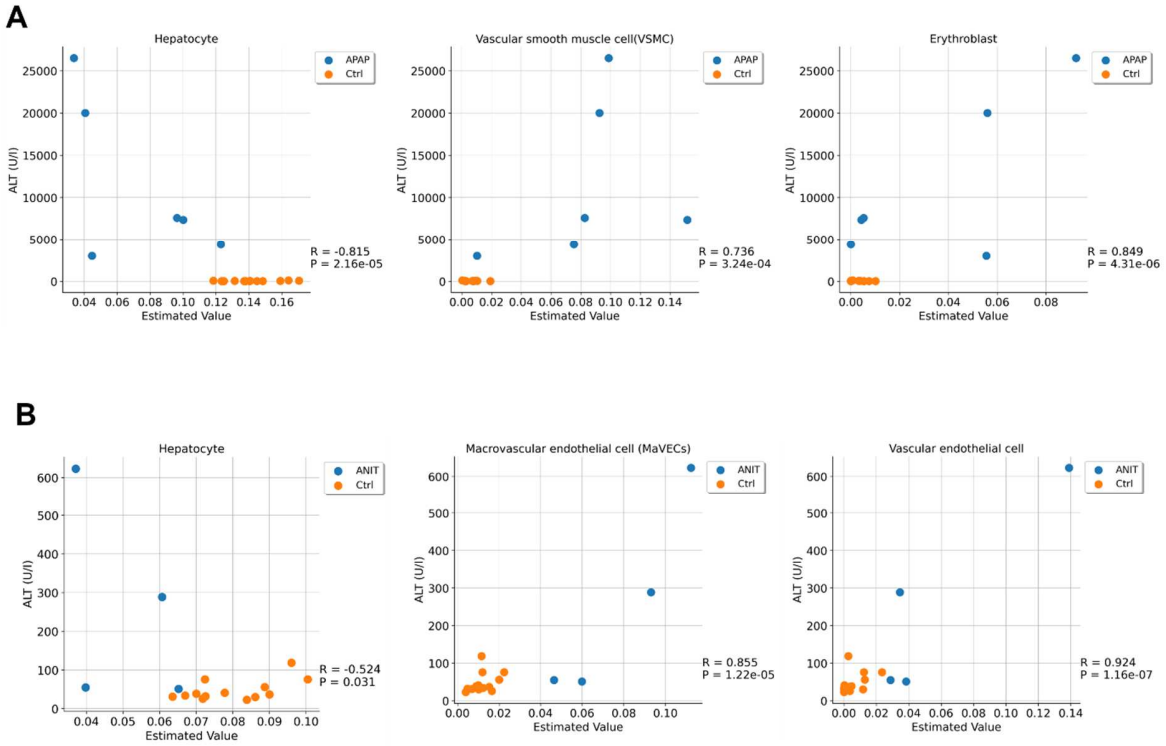


Figure S6. Relationship between the cell ratio estimated by GLDADec and the degree of injury. Scatterplots comparing estimated proportions and blood biochemistry value in **(A)** APAP and **(B)** ANIT treatment group, respectively. Alanine aminotransferase (ALT) is the most common marker of liver damage. High ALT values are a sign of severe liver damage.

Figure S7

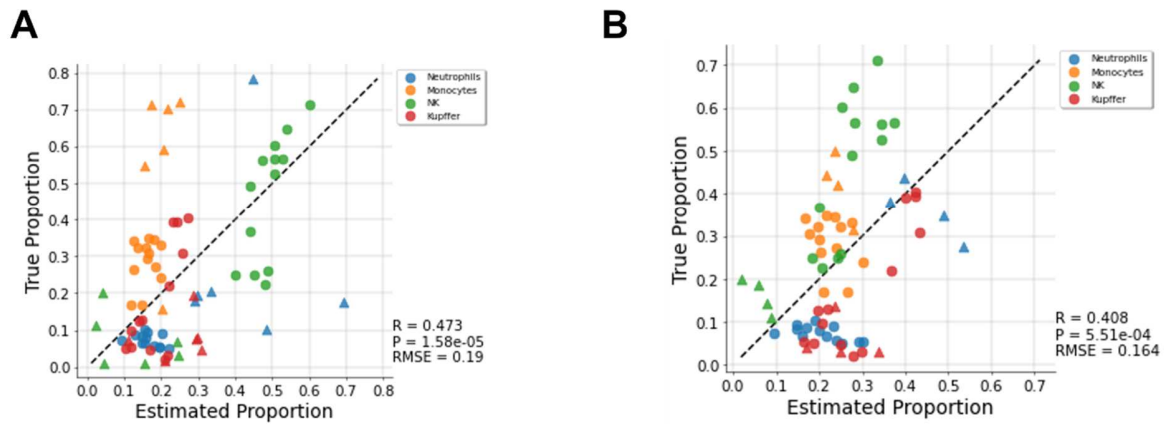


Figure S7. Estimation performance of GLDADec when using differentially expressed genes (DEGs) derived marker genes as prior information. Scatterplot showing the estimated proportions of immune cells vs. measured values for **(A)** acetaminophen and **(B)** alpha-naphthyl isothiocyanate administration group, respectively.

Figure S8

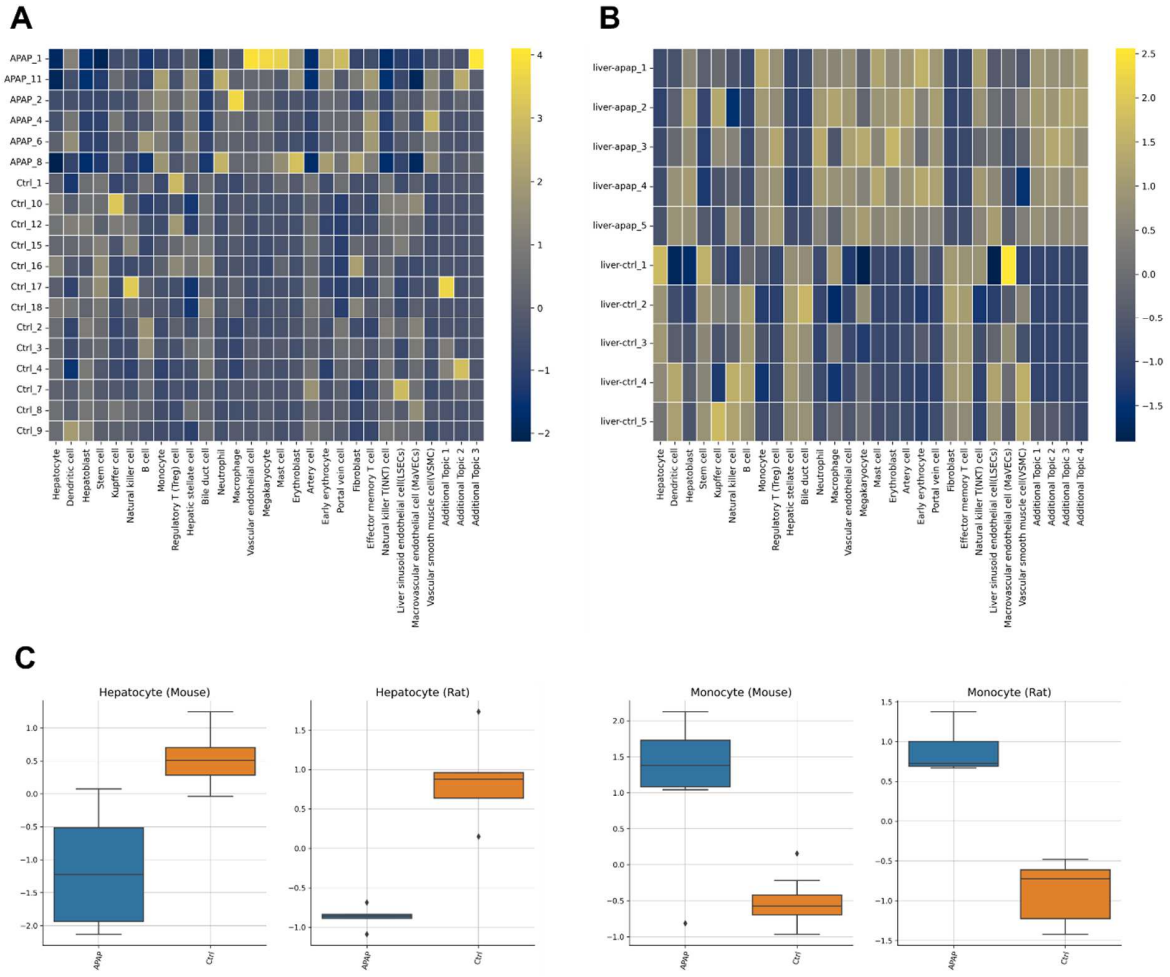


Figure S8. Comparison of immune cell trafficking patterns in mice and rats after acetaminophen (APAP) administration. Heatmaps showing the estimated comprehensive cell type proportions in **(A)** mice and **(B)** rats. **(C)** Common cell trafficking patterns between mice and rats.

Figure S9

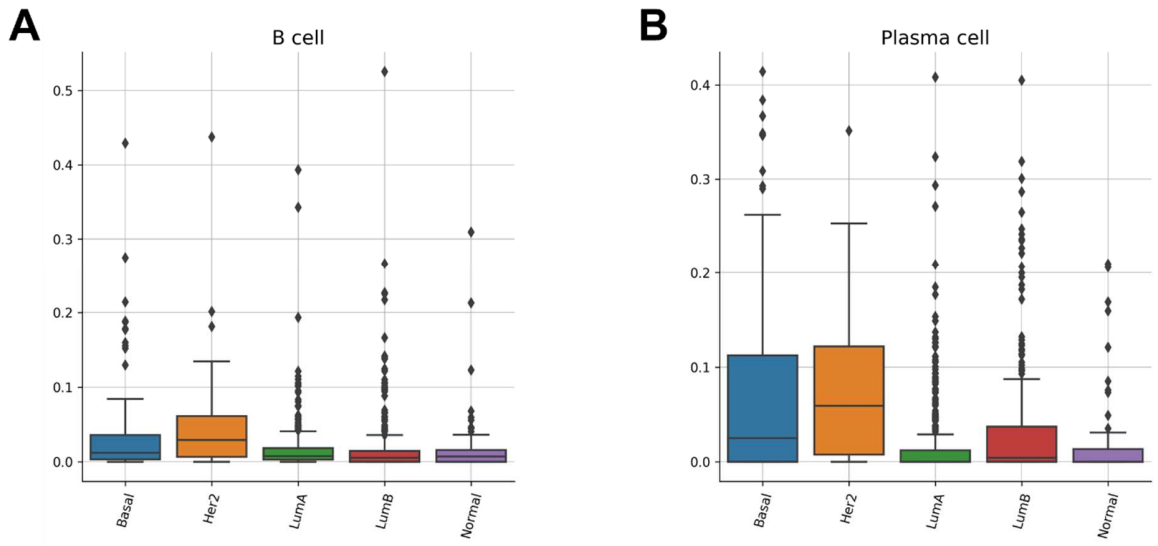


Figure S9. Comparison of cells showing specific accumulation and infiltration between each subtype of BRCA. Her2 subtype-specific accumulation was observed in **(A)** B cells and **(B)** plasma cells, respectively.

Figure S10

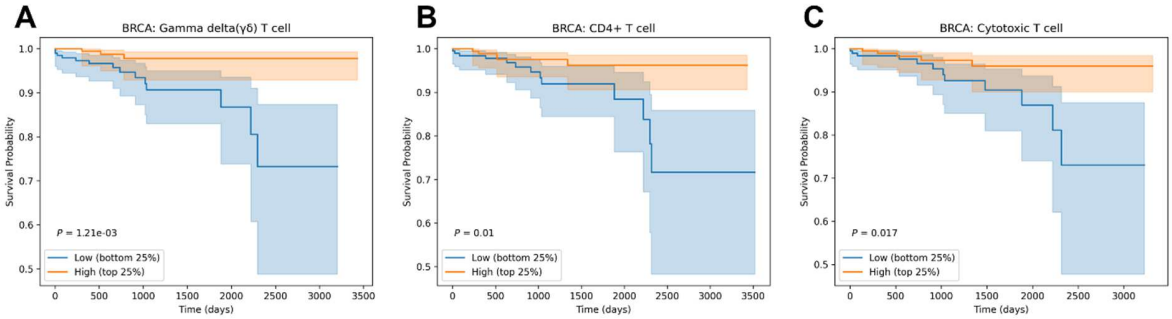


Figure S10. Survival analysis when samples are stratified by the accumulation of cells estimated by GLDADec. Kaplan-Meier plots showing survival associations with infiltration of **(A)** gamma delta (γδT) cells, **(B)** CD4+ T cells, and **(C)** Cytotoxic T cells in BRCA. Patients with the top 25th percentile of the target cells were compared with those with the bottom 25th percentile.

Figure S11

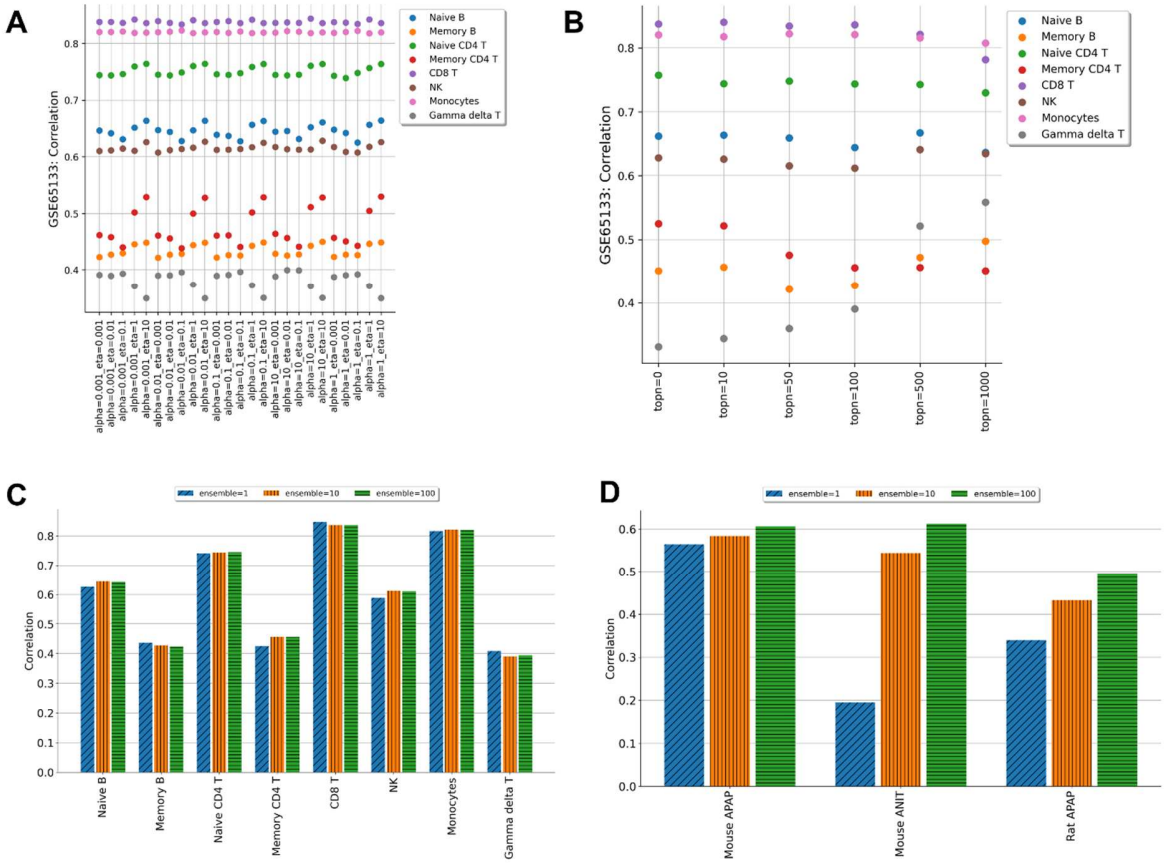


Figure S11. Hyperparameter sensitivity analysis. **(A)** Effect of the combination of hyperparameters α and η on the Dirichlet distributions. **(B)** Relationship between the number of additional genes with large coefficients of variation and estimation performance. Bar plots representing the contribution of the ensemble to improved estimation performance for **(C)** blood-derived samples and **(D)** tissue samples, respectively.

Supplementary tables

Table S1

Table S1. Summary of datasets used in this study.

Dataset	PMID	Size	Species	Source
GSE65133	25822800	20	Human	PBMC
GSE107572	31126321	9	Human	PBMC/PMN
GSE60424	25314013	5	Human	Whole blood
GSE237801-APAP	38187088	19	Mouse	Liver
GSE237801-ANIT	38187088	17	Mouse	Liver
GSE239996-APAP	37941435	10	Rat	Liver
TCGA-BRCA	-	1052	Human	Breast
TCGA-LUAD	-	577	Human	Lung
TCGA-LIHC	-	408	Human	Liver

Table S2

Table S2. Gene ontology (GO) enrichment analysis. Significantly enriched GO terms are shown with Bonferroni adjusted *P*-values.

Topics	Gene Ontology	Adjusted *P*-value	Overlap
2	Lipid Biosynthetic Process (GO:0008610)	0.003751	{'FITM1', 'FDFT1', 'ACSM3', 'GPAT4', 'ACACA', 'SRD5A1', 'ACSM5', 'PRLR', 'ACACB'}
	Fatty Acid Metabolic Process (GO:0006631)	0.004734	{'PNPLA3', 'ACSM3', 'PPARD', 'ACADSB', 'PCCA', 'FAAH', 'HPGD', 'ACACA', 'ACACB', 'ACSM5', 'ACSL4'}
	Monocarboxylic Acid Biosynthetic Process (GO:0072330)	0.04295	{'SRR', 'ACSM3', 'CYP7B1', 'AKR1D1', 'ACACA', 'ACSM5', 'ACACB'}
3	Response To Lipopolysaccharide (GO:0032496)	0.00002702	{'PF4', 'PDE4B', 'CXCL3', 'SELE', 'ZC3H12A', 'NFKBIB', 'TNIP1', 'TNFAIP3', 'TNIP2', 'SELP', 'CXCL13', 'IL1A', 'SBNO2', 'MAPKAPK2', 'IRAK3', 'CD274'}
	Cellular Response To Lipopolysaccharide (GO:0071222)	0.0001060	{'PDE4B', 'CXCL3', 'CCL3', 'ZC3H12A', 'NFKBIB', 'TNIP1', 'TNFAIP3', 'TNIP2', 'CXCL13', 'IL1A', 'SBNO2', 'PF4', 'CD274'}
	Response To Interleukin-1 (GO:0070555)	0.0001060	{'IRAK2', 'CCL7', 'CCL3', 'SOX9', 'SELE', 'SRC', 'ZC3H12A', 'HIF1A', 'CCL4', 'IRAK3', 'CCL19'}
	Response To Cytokine (GO:0034097)	0.0001060	{'IRAK2', 'SPHK1', 'SELE', 'SRC', 'CSF3', 'TIMP1', 'TIMP3', 'SELP', 'IFITM1', 'CH25H', 'MAPKAPK2', 'IRAK3', 'CD274'}
	Cellular Response To Molecule Of Bacterial Origin (GO:0071219)	0.0002890	{'PDE4B', 'CXCL3', 'ZC3H12A', 'NFKBIB', 'TNIP1', 'TNFAIP3', 'TNIP2', 'CXCL13', 'IL1A', 'SBNO2', 'PF4', 'CD274'}
	Cellular Response To Interleukin-1 (GO:0071347)	0.0003670	{'IRAK2', 'CCL7', 'CCL3', 'SOX9', 'ZC3H12A', 'TNIP2', 'HIF1A', 'CCL4', 'IRAK3', 'CCL19'}
	Cellular Response To Lipid (GO:0071396)	0.0005460	{'PDE4B', 'CXCL3', 'SOX9', 'PDK4', 'ZC3H12A', 'NFKBIB', 'HSPA1B', 'TNIP1', 'TNFAIP3', 'TNIP2', 'CXCL13', 'IL1A', 'SBNO2', 'HSPA1A', 'PF4', 'CD274'}
	Cellular Response To Chemokine (GO:1990869)	0.001605	{'CCL7', 'CXCL3', 'CCL3', 'ZC3H12A', 'CXCL13', 'CCL4', 'PF4', 'CCL19'}
	Cytokine-Mediated Signaling Pathway (GO:0019221)	0.001970	{'IRAK2', 'SOCS1', 'CCL7', 'PF4', 'CCL3', 'CXCL3', 'SRC', 'CSF3', 'LEPR', 'TNIP2', 'CXCL13', 'IL1A', 'CCL4', 'IRAK3', 'CCL19', 'IL17RA'}
	Regulation Of Tumor Necrosis Factor-Mediated Signaling Pathway (GO:0010803)	0.002976	{'SPHK1', 'HSPA1A', 'PTPN2', 'HSPA1B', 'PPP2CB', 'TNFAIP3', 'MAPKAPK2'}
	Neutrophil Chemotaxis	0.004221	{'PDE4B', 'CCL7', 'CCL3', 'CXCL3', 'CXCL13', 'CCL4',

	(GO:0030593)		'PF4', 'CCL19'}
	Granulocyte Chemotaxis (GO:0071621)	0.005270	{'PDE4B', 'CCL7', 'CCL3', 'CXCL3', 'CXCL13', 'CCL4', 'PF4', 'CCL19'}
	Neutrophil Migration (GO:1990266)	0.006068	{'PDE4B', 'CCL7', 'CCL3', 'CXCL3', 'CCL19', 'CCL4', 'PF4', 'CXCL13'}
	Chemokine-Mediated Signaling Pathway (GO:0070098)	0.006068	{'CCL7', 'CXCL3', 'CCL3', 'CXCL13', 'CCL4', 'PF4', 'CCL19'}
	Regulation Of Cell Population Proliferation (GO:0042127)	0.006068	{'KLF10', 'FOSL1', 'SOX9', 'ADAMTS1', 'PTPN2', 'CSF3', 'TIMP1', 'NGF', 'TBRG1', 'BTG3', 'SGK1', 'GML', 'GCNT2', 'IL1A', 'BTG2', 'PF4', 'IER5', 'CSF1', 'TNFRSF9', 'HILPDA', 'IGFBP3', 'FGF23', 'HSPA1A', 'MYC', 'HSPA1B', 'CD200', 'RUNX1', 'IFITM1', 'TES'}
	Cellular Response To Heat (GO:0034605)	0.006068	{'HSPA1B', 'BAG3', 'IL1A', 'DNAJB1', 'HSPA1A', 'IER5'}
	Regulation Of Cytokine-Mediated Signaling Pathway (GO:0001959)	0.006068	{'IRAK2', 'SPHK1', 'PTPN2', 'TNFAIP3', 'MAPKAPK2', 'IRAK3'}
	Regulation Of Natural Killer Cell Chemotaxis (GO:2000501)	0.009188	{'CCL4', 'CCL7', 'CCL3'}
	Lymphocyte Chemotaxis (GO:0048247)	0.009188	{'CCL7', 'CCL3', 'CXCL13', 'CCL4', 'CH25H', 'CCL19'}
	Negative Regulation Of Cell Population Proliferation (GO:0008285)	0.009188	{'KLF10', 'IGFBP3', 'BTG3', 'SOX9', 'ADAMTS1', 'PTPN2', 'HSPA1B', 'CD200', 'GML', 'TNFAIP3', 'NGF', 'IL1A', 'BTG2', 'IFITM1', 'HSPA1A', 'TBRG1', 'TES', 'MYC'}
	Positive Regulation Of Lymphocyte Chemotaxis (GO:0140131)	0.01459	{'CCL4', 'CCL7', 'CCL3', 'CXCL13'}
	Negative Regulation Of Extrinsic Apoptotic Signaling Pathway (GO:2001237)	0.01597	{'SRC', 'HSPA1B', 'TNFAIP3', 'IL1A', 'ICAM1', 'HSPA1A', 'PF4'}
	Myeloid Leukocyte Differentiation (GO:0002573)	0.01670	{'CSF1', 'SOCS1', 'CSF3', 'RUNX1', 'SBNO2', 'EPHA2', 'CCL19'}
	Cellular Response To Oxygen-Containing Compound (GO:1901701)	0.01797	{'PDE4B', 'CXCL3', 'SOX9', 'PDK4', 'SLC13A5', 'ZC3H12A', 'NFKBIB', 'LCE1D', 'TNIP1', 'TNFAIP3', 'TNIP2', 'P2RY2', 'CXCL13', 'IL1A', 'CPNE8', 'SBNO2', 'PF4', 'CD274'}
	Proline Transport (GO:0015824)	0.01897	{'SLC3A2', 'SLC7A5', 'SLC38A2'}
	Cellular Response To Cytokine Stimulus (GO:0071345)	0.02046	{'ZFP36L2', 'SOCS1', 'CCL7', 'PF4', 'CCL3', 'SOX9', 'CSF3', 'ZC3H12A', 'LEPR', 'CCL4', 'IL1A', 'HIF1A', 'SBNO2', 'IRAK3', 'CCL19'}
	Leucine Transport (GO:0015820)	0.02516	{'SLC3A2', 'SLC7A5', 'SLC7A6'}

	Negative Regulation Of Biomineral Tissue Development (GO:0070168)	0.02516	{'FGF23', 'CCL3', 'SRGN', 'SOX9'}
	Regulation Of Anoikis (GO:2000209)	0.02516	{'SRC', 'SIK1', 'MYBBP1A', 'PDK4'}
	Negative Regulation Of NF-kappaB Transcription Factor Activity (GO:0032088)	0.02749	{'IRAK2', 'ZC3H12A', 'NFKBIB', 'CD200', 'PPP2CB', 'TNFAIP3', 'IRAK3'}
	Negative Regulation Of Extrinsic Apoptotic Signaling Pathway In Absence Of Ligand (GO:2001240)	0.02749	{'HSPA1A', 'PF4', 'HSPA1B', 'IL1A'}
	Negative Regulation Of Signal Transduction In Absence Of Ligand (GO:1901099)	0.02749	{'HSPA1A', 'PF4', 'HSPA1B', 'IL1A'}
	Response To Tumor Necrosis Factor (GO:0034612)	0.02751	{'ZFP36L2', 'SPHK1', 'CCL7', 'CCL3', 'SELE', 'ZC3H12A', 'CCL4', 'CCL19'}
	Negative Regulation Of Inclusion Body Assembly (GO:0090084)	0.02751	{'DNAJB1', 'HSPA1A', 'HSPA1B'}
	Branched-Chain Amino Acid Transport (GO:0015803)	0.02751	{'SLC3A2', 'SLC7A5', 'SLC7A6'}
	Regulation Of Inclusion Body Assembly (GO:0090083)	0.02751	{'DNAJB1', 'HSPA1A', 'HSPA1B'}
	Negative Regulation Of Ossification (GO:0030279)	0.03273	{'FGF23', 'CCL3', 'SRGN', 'SOX9'}
	ERK1 And ERK2 Cascade (GO:0070371)	0.03273	{'ZFP36L2', 'DUSP5', 'MYC', 'SOX9'}
	Biogenic Amine Biosynthetic Process (GO:0042401)	0.03448	{'SMOX', 'HDC', 'SLC38A2'}
	Inflammatory Response (GO:0006954)	0.03563	{'IRAK2', 'CCL7', 'CXCL3', 'CCL3', 'SELE', 'ACKR1', 'CXCL13', 'IL1A', 'CCL4', 'MAPKAPK2', 'PF4', 'CCL19'}
	Osteoclast Differentiation (GO:0030316)	0.03563	{'CSF1', 'SRC', 'EPHA2', 'SBNO2'}
	Regulation Of Angiogenesis (GO:0045765)	0.03675	{'SPHK1', 'NPPB', 'ADAMTS1', 'ZC3H12A', 'TNFAIP3', 'RUNX1', 'IL1A', 'HIF1A', 'PF4', 'EPHA2', 'CXCL13'}
	MyD88-dependent Toll-Like Receptor Signaling Pathway (GO:0002755)	0.04117	{'IRAK2', 'TNIP1', 'IRAK3'}
	Negative Regulation Of Cellular Process (GO:0048523)	0.04155	{'KLF10', 'ADAMTS1', 'PTPN2', 'TSPYL2', 'NGF', 'TBRG1', 'CDA', 'BTG3', 'GML', 'IL1A', 'BTG2', 'NPPB', 'NR1H2', 'IGFBP3', 'HSPA1A', 'MYC', 'HSPA1B', 'CD200', 'IFITM1', 'TES'}
	Regulation Of Extrinsic Apoptotic	0.04314	{'HSPA1A', 'PF4', 'HSPA1B', 'IL1A'}

	Signaling Pathway In Absence Of Ligand (GO:2001239)		
	Negative Regulation Of DNA-binding Transcription Factor Activity (GO:0043433)	0.04314	{'IRAK2', 'BHLHE40', 'ZC3H12A', 'CD200', 'PPP2CB', 'TNFAIP3', 'SIK1', 'PIM1', 'IRAK3'}
	Negative Regulation Of Bone Mineralization (GO:0030502)	0.04834	{'FGF23', 'CCL3', 'SRGN'}
4	Lipid Biosynthetic Process (GO:0008610)	0.006111	{'FDFT1', 'MVD', 'GPAT4', 'ACACA', 'SRD5A1', 'MVK', 'ACACB', 'CYP17A1'}
	Secondary Alcohol Biosynthetic Process (GO:1902653)	0.006361	{'FDFT1', 'MVD', 'TM7SF2', 'MVK', 'PMVK'}
	Cholesterol Biosynthetic Process (GO:0006695)	0.006361	{'FDFT1', 'MVD', 'TM7SF2', 'MVK', 'PMVK'}
	Sterol Biosynthetic Process (GO:0016126)	0.008519	{'FDFT1', 'MVD', 'TM7SF2', 'MVK', 'PMVK'}
	acetyl-CoA Metabolic Process (GO:0006084)	0.01020	{'MVK', 'PMVK', 'ACACB', 'MVD'}

Table S3

Table S3. Pearson correlation between cell proportions estimated by GLDADec and various liver injury markers such as ALT, AST, and TBIL after acetaminophen (APAP) administration.

APAP			
	ALT	AST	TBIL
ALT	1.0000	0.9478	0.8273
Neutrophil	0.9776	0.9113	0.8409
AST	0.9478	1.0000	0.8868
Erythroblast	0.8489	0.8944	0.8514
Monocyte	0.8422	0.6611	0.5730
TBIL	0.8273	0.8868	1.0000
Early erythrocyte	0.7494	0.6764	0.6854
Vascular smooth muscle cell(VSMC)	0.7364	0.5258	0.4420
Effector memory T cell	0.6934	0.4840	0.4238
Fibroblast	0.5214	0.5888	0.6284
Hepatic stellate cell	0.4985	0.3102	0.2644
Additional Topic 2	0.3274	0.2759	0.1838
Macrophage	0.3213	0.1122	0.0326
Mast cell	0.3049	0.3649	0.4674
Megakaryocyte	0.3037	0.2296	0.3121
Portal vein cell	0.2071	0.1524	0.1219
Vascular endothelial cell	0.2039	0.1528	0.2393
Kupffer cell	0.1402	0.1229	-0.0461
Additional Topic 3	0.0945	0.0924	0.2377
Dendritic cell	-0.1666	-0.2380	-0.2165
Regulatory T (Treg) cell	-0.1757	-0.1322	-0.2198
Additional Topic 1	-0.1878	-0.1354	0.0641
Liver sinusoid endothelial cell(LSECs)	-0.2931	-0.3024	-0.3350
Natural killer cell	-0.3025	-0.2645	-0.0671
B cell	-0.3654	-0.4351	-0.5158
Stem cell	-0.6069	-0.4646	-0.3353
Bile duct cell	-0.6770	-0.5030	-0.4584
Macrovascular endothelial cell (MaVECs)	-0.7418	-0.6709	-0.7796
Artery cell	-0.7590	-0.6176	-0.5497
Hepatoblast	-0.7598	-0.6140	-0.6680

Natural killer T(NKT) cell	-0.7752	-0.6146	-0.6192
Hepatocyte	-0.8147	-0.7338	-0.7379

Table S4

Table S4. Pearson correlation between cell proportions estimated by GLDADec and various liver injury markers such as ALT, AST, and TBIL after alpha-naphthyl isothiocyanate (ANIT) administration.

ANIT			
	ALT	AST	TBIL
ALT	1.0000	0.9837	0.8807
AST	0.9837	1.0000	0.9075
Vascular endothelial cell	0.9242	0.9505	0.9681
TBIL	0.8807	0.9075	1.0000
Macrovascular endothelial cell (MaVECs)	0.8551	0.8211	0.8922
Artery cell	0.6627	0.6265	0.8123
Early erythrocyte	0.6430	0.5935	0.6668
Neutrophil	0.5313	0.5681	0.6912
Hepatoblast	0.4747	0.4362	0.6652
Additional Topic 1	0.4442	0.4637	0.5498
Portal vein cell	0.3769	0.4103	0.4922
Monocyte	0.3247	0.3493	0.4534
Mast cell	-0.0027	0.0103	0.1332
Natural killer T(NKT) cell	-0.0639	-0.1112	-0.2240
Megakaryocyte	-0.0988	-0.1067	-0.0829
Natural killer cell	-0.1917	-0.2264	-0.2086
Additional Topic 3	-0.2497	-0.2274	-0.3192
Regulatory T (Treg) cell	-0.2708	-0.2443	-0.2546
Erythroblast	-0.2835	-0.2273	-0.2528
Kupffer cell	-0.3059	-0.3291	-0.4473
Additional Tpic 2	-0.3316	-0.3029	-0.4250
Macrophage	-0.3538	-0.3653	-0.4317
Stem cell	-0.3539	-0.4019	-0.5307
Vascular smooth muscle cell(VSMC)	-0.3867	-0.3159	-0.3996
Effector memory T cell	-0.4318	-0.4064	-0.4854
Fibroblast	-0.4594	-0.4849	-0.5025

Hepatic stellate cell	-0.4704	-0.4659	-0.5271
Bile duct cell	-0.4939	-0.5017	-0.6166
Liver sinusoid endothelial cell(LSECs)	-0.5002	-0.4884	-0.6263
Hepatocyte	-0.5237	-0.5715	-0.6563
Dendritic cell	-0.5889	-0.5980	-0.7413
B cell	-0.6192	-0.5832	-0.7398

Supplementary notes

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

FUNDING

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (C) (grant number 21K06663) from the Japan Society for the Promotion of Science, Takeda Science Foundation, and Mochida Memorial Foundation for Medical and Pharmaceutical Research.

Note S1. Data processing

In this section, we will outline the detailed data processing methods. The processed and input data can be accessed on our Github repository (https://github.com/mizuno-group/GLDADec).

Blood-derived benchmarking data

The NCBI Gene Expression Omnibus (GEO) database was the source of the three datasets, GSE65133, GSE107572, and GSE60424, which were obtained as processed series matrix files [1–3]. These datasets were derived from human samples, and the transcript IDs were converted to HGNC symbols using files available from Biomart [4]. Only immune cell types measured by flow cytometry and included in the benchmark data were selected for evaluation.

Tissue data with perturbation

In this study, we utilized bulk RNA-Seq data from liver tissue samples obtained from drug-induced liver injury models. The mouse and rat datasets were obtained from our previous studies, which are publicly available in GEO under accession numbers GSE237801 and GSE239996, respectively [5,6]. Transcripts per kilobase million (TPM) values were obtained for the data, and the transcript IDs were converted to gene symbols. The median values were selected for genes with duplicated names. The control samples and the samples of interest, such as the acetaminophen-treated group, were merged and used as targets to estimate the cellular changes associated with the perturbation. To highlight the differences in the trafficking of immune cells due to perturbation, a feature-wise minmax correction was performed as a preprocessing step to emphasize differences in expression levels between samples. Cell types verified by flow cytometry were evaluated for deconvolution, including neutrophils, monocytes, NK cells, and Kupffer cells in mice, and B cells, CD4+ T cells, CD8 + T cells, neutrophils, monocytes, and NK cells in rats.

Clinical data

The TCGA RNA-seq data was obtained on November 16, 2022. The datasets were then prepared for three types of tumors, namely breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), and liver hepatocellular carcinoma (LIHC), and were normalized using TPM [7–10]. To mitigate the impact of the distribution of expression levels in each sample on the estimated values, quantile normalization was performed. Due to the large number of samples and the high computational demands of these datasets, the linear scale expression was transformed into a sparser matrix by dividing it by a constant value.

Note S2. Condition of experiments

In this section, we provide a comprehensive description of the execution environment of the proposed and existing methods in various scenarios, which can be replicated in our GitHub repository (https://github.com/mizuno-group/GLDADec).

Benchmarking with human blood samples

We employed a linear scaled blood-derived benchmark dataset, utilizing marker gene names defined by domain knowledge (see Supplementary File S1) as partial prior information. Genes exhibiting outliers greater than 2σ from the log-normal distribution of gene expression levels were eliminated for each sample, and 100 genes displaying substantial sample-wide coefficients of variation (CV) were incorporated into the analysis. This number of genes to be added serves as a hyperparameter and is further discussed in the main text. It is important to note that the benchmark data originate from blood samples, and no additional topics were integrated into the model since the constituent cell types are clearly delineated. To correct for the expression level on a linear scale, the values were divided by constants specific to each dataset, i.e., 10, 200, and 100 for GSE6513, GSE107572, and GSE60424, respectively. This operation was implemented to save computational resources and does not alter the relative gene expression levels of each sample.

We conducted a comparison of the performance of GLDADec with seven alternative deconvolution methods, including FARDEEP, EPIC, CIBERSORTx, DCQ, NNLS, RLR, and Elastic Net [11–14]. These methods are all reference-based and require prior knowledge of a gene signature matrix as the reference. The LM22 definition provided by Newman et al. is commonly used as a reference for human blood-derived samples and was also obtained from their work and utilized in our analysis [1]. Notably, LM22 was defined in the same study that generated GSE65133, one of the benchmark datasets used in our analysis, making it a compatible reference. The competing methods have diverse development backgrounds and accept different formats of expression data as input. We adhered to their default settings whenever possible. Specifically, we employed log-normalized gene expression matrices for DCQ and FARDEEP, and a non-log linear scaled matrix for the remaining methods. NNLS, RLR, and Elastic Net did not specify the input format, so we adapted them to match the scale of the reference data. To eliminate the influence of differences in scale between the analysis object and the reference on the results, we converted LM22 to a log-scale in the DCQ and FARDEEP analysis.

Comprehensive cell type analysis for mouse perturbed tissue data

The dataset GSE237801 comprises bulk RNA-Seq data from mouse liver tissue under various compound administration and immune cell proportions measured by flow cytometry, which is a suitable benchmark dataset for deconvolution methods for tissues [6]. This study primarily focused on the acetaminophen (APAP) and alpha-naphthyl isothiocyanate (ANIT) treatment groups, examining the performance of deconvolution methods in detecting perturbation-induced changes in immune cell proportions relative to control samples. It is important to note that the tissue sample is a heterogeneous cell population, and subsequent analyses incorporate additional modeling considerations.

In our analysis of benchmarking using human blood datasets, GLDADec utilized linear scale expression levels as input and marker genes as prior information. In the examination of tissues, it is anticipated that transcriptome variation will arise due to the presence of unknown cells and confounding factors. To address this, we gathered liver-related cell types and their associated marker genes in a data-driven manner from the CellMarker database [15]. Moreover, to account for the diverse factors affecting the tissue transcriptome, we included 1000 genes with substantial coefficients of variation, which adequately describe the state of the tissue.

We contrasted the performance of GLDADec with seven existing reference-based methods, which are identical to those analyzed using the blood-derived benchmark dataset. Since the LM22 expression levels are derived from human samples, we employed the mouse-derived LM13, as previously defined in our analysis, as a reference to account for species differences in the mouse data. Additionally, we performed an analysis in which differentially expressed gene (DEGs) names derived from the same LM13 were considered as markers and subjected to GLDADec.

Application to rat data

GSE239996 contains data on rat liver tissue bulk RNA-Seq under various compound administration and immune cell proportions measured by flow cytometry [5]. In this study, we concentrated on the acetaminophen (APAP) treatment groups and assessed the efficacy of the deconvolution method in detecting immune cell proportion alterations induced by the compound compared to control samples. Similar to the analysis of the mouse data, 1000 genes with significant expression variation were included.

In this section, we defined three types of marker gene names for GLDADec analysis. The first set of classical cell type markers were collected from Figure 4A of Natasha's work [16]. The second marker is the same as defined in the aforementioned comprehensive analysis of mice from CellMarker. For the last set, DEGs were calculated for the six cell types (LM6) out of LM13, and the corresponding gene

names were designated as markers. Utilizing LM6 as a reference, we compared its performance with seven competing methods.

Application to tumor samples

The TCGA RNA-seq data were obtained from numerous experiments and consist of a large number of samples, but also various confounding factors. To address this, we performed quantile normalization and aligned the distribution of each sample. However, the policy of eliminating genes with values that are more than two standard deviations away from the mean did not ensure enough genes, so we instead included genes with large expression variation in this analysis.

We gathered marker gene names from the background tissues of each cancer subtype using CellMarker, and utilized this information as prior knowledge. Specifically, we obtained markers for 47, 73, and 80 cell types in a data-driven manner for breast, lung, and liver cancers, respectively. It is important to note that these markers were derived from human clinical data.

References

1. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. Nature Methods 2015; 12:453–457

2. Finotello F, Mayer C, Plattner C, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Medicine 2019; 11:34

3. Linsley PS, Speake C, Whalen E, et al. Copy Number Loss of the Interferon Gene Cluster in Melanomas Is Linked to Reduced T Cell Infiltrate and Poor Patient Prognosis. PLoS ONE 2014; 9:e109760

4. Smedley D, Haider S, Ballester B, et al. BioMart – biological queries made easy. BMC Genomics 2009; 10:22

5. Morita K, Mizuno T, Azuma I, et al. Rat deconvolution as knowledge miner for immune cell trafficking from toxicogenomics databases. Toxicological Sciences 2023;

6. Azuma I, Mizuno T, Morita K, et al. Investigation of the usefulness of liver-specific deconvolution method by establishing a liver benchmark dataset. NAR Genom Bioinform 2024; 6:

7. Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics 2013; 45:1113–1120

8. Berger AC, Korkut A, Kanchi RS, et al. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. Cancer Cell 2018; 33:690-705.e9

9. Campbell JD, Alexandrov A, Kim J, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nature Genetics 2016; 48:607–616

10. Ally A, Balasundaram M, Carlsen R, et al. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell 2017; 169:1327-1341.e23

11. Hao Y, Yan M, Heath BR, et al. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. PLOS Computational Biology 2019; 15:e1006976

12. Racle J, de Jonge K, Baumgaertner P, et al. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. eLife 2017; 6:

13. Steen CB, Liu CL, Alizadeh AA, et al. Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx. Methods in molecular biology (Clifton, N.J.) 2020; 2117:135–157

14. Altboum Z, Steuerman Y, David E, et al. Digital cell quantification identifies global immune cell dynamics during influenza infection. Molecular Systems Biology 2014; 10:

15. Hu C, Li T, Xu Y, et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids Research 2023; 51:D870–D876

16. Qaisar N, Arowosegbe A, Derr AG, et al. Type I IFN–Driven Immune Cell Dysregulation in Rat Autoimmune Diabetes. ImmunoHorizons 2021; 5:855–869