DWEN: A novel method for accurate estimation of cell type compositions from bulk data samples

Duc Tran
Computer Science & Engineering
University of Nevada, Reno
Reno, USA
duct@nevada.unr.edu

Ha Nguyen
Computer Science & Engineering
University of Nevada, Reno
Reno, USA
hanguyen@nevada.unr.edu

Hung Nguyen
Computer Science & Engineering
University of Nevada, Reno
Reno, USA
hungnp@nevada.unr.edu

Tin Nguyen*
Computer Science & Engineering
University of Nevada, Reno
Reno, USA
tinn@unr.edu

2022 14th International Conference on Knowledge and Systems Engineering (KSE) | 978-1-6654-5281-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/KSE56063.2022.9953757

***Abstract*—Advances in single-cell RNA sequencing (scRNA-seq) technologies have allowed us to study the heterogeneity of cell populations. The cell compositions of tissues from different hosts may vary greatly, indicating the condition of the hosts, from which the samples are collected. However, the high sequencing cost and the lack of fresh tissues make single-cell approaches less appealing. In many cases, it is practically impossible to generate single-cell data in a large number of subjects, making it challenging to monitor changes in cell type compositions in various diseases. Here we introduce a novel approach, named Deconvolution using Weighted Elastic Net (DWEN), that allows researchers to accurately estimate the cell type compositions from bulk data samples without the need of generating single-cell data. It also allows for the re-analysis of bulk data collected from rare conditions to extract more in-depth cell-type level insights. The approach consists of two modules. The first module constructs the cell type signature matrix from single-cell data while the second module estimates the cell type compositions of input bulk samples. In an extensive analysis using 20 datasets generated from scRNA-seq data of different human tissues, we demonstrate that DWEN outperforms current state-of-the-arts in estimating cell type compositions of bulk samples.**

***Index Terms*—scRNA-seq, bulk data deconvolution**

I. Introduction

Complex biological tissues consist of multiple cell types with varying proportions. Cell type proportions may play central roles in controlling host responses to physiological and pathological conditions [1]. Studying the cell compositions of these tissues under various conditions provides valuable insights into mechanisms of underlying diseases. For example, the composition of immune cells in tumor micro-environments is one of the main contributors to cancer's heterogeneity [2]. It has been demonstrated that immune cells infiltrate tumors to regulate their growth and its composition within the solid tumor is an important indicator of patients' survival [3].

The cell type compositions of tissues can be investigated via laboratory methods like flow cytometry [4], laser capture micro-dissection [5], and immunohistochemistry [6]. Alternatively, single cell RNA sequencing (scRNA-seq) technologies have provided a powerful approach to systematically capture the cellular heterogeneity and identify new cell types [7]. However, scRNA-seq data still have critical limitations, including: (i) high sequencing cost, (ii) technical noise, and (iii) inappropriate reflection of the cell type proportions of the tissue [8]. To overcome these limitations, deconvolution methods have been developed to obtain the constitution of cells directly from bulk expression data. This enables the ability to infer cell type proportions in bulk tissues, thus allows us to study cell heterogeneity within tissues without the need to dissolve the bulk samples into individual cells.

Current deconvolution methods methods can be classified into two main categories, reference-free and reference-based methods. Methods in the first category include BayCount [9], BayesCCE [10], CellDistinguisher [11], deconf [12], and TOAST [13]. These methods usually rely on matrix factorization or statistical methods to decompose the input data into the signature matrix and the corresponding cell type proportions. The disadvantage of reference-free methods includes: (i) the exact cell types correspond to the inferred proportions are unknown, and (ii) heavy computation. Methods in the second category include AdRoit [14], DAISM [15], EPIC [16], quanTIseq [17], and SCDC [18]. These methods usually come with a pipeline following two sequential steps: (i) constructing a signature for each cell type and (ii) deconvolve the bulk data using the obtained signature matrix. One drawback of the reference-based methods is their estimations often have biases against the cell types with lower proportions, or cell types characterized by markers with low expressions.

Here we propose a novel approach, Deconvolution using Weighted Elastic Net (DWEN), that can accurately infer cell type compositions of bulk samples. The novel *weighted elastic net* approach has many advantages over current methods: (i) robustness to noise as the weights prevent the model from only focusing on reducing the residual errors in the highly expressed genes and ignoring the rest; (ii) automatic and reliable feature selection from the signature matrix; and (iii) better

prediction performance as tolerance to multi-collinearity. We demonstrate that DWEN outperforms state-of-the-art methods and accurately estimates the true cell type proportions in 20 datasets obtained from the human cell atlas [19].

II. Method

Figure 1 depicts the overall workflow of DWEN. The method requires the following input: (i) a reference single-cell expression dataset of known cell types, and (ii) a bulk gene expression dataset that needs to be deconvolved. The first input includes a matrix of genes by cells and a vector indicating the cell type label of each cell. The second input, the bulk expression dataset, is a matrix of genes by samples.

Given the bulk expression and the reference single-cell data, the method follows a pipeline of two sequential modules: (i) constructing cell type signature matrix from the single-cell data, and (ii) estimating the proportion of each cell type in the bulk samples using a linear regression model named *weighted elastic net*. The details are described in the following sections.

A. Module I: Signature matrix construction

The goal of this module is to construct a signature matrix of known cell types and their biomarkers. First, we perform a gene filtering step to reduce noise in the reference single-cell data. For each cell type, we select a set of genes that are expressed in at least 30% of the cells of the cell type. Next, we remove genes that are not presented in any of those gene sets obtained from the previous step. This allows us to keep genes that have sparse expression profiles in the whole scRNA-seq expression matrix, but not in a particular cell type. At the same time, we also remove genes that might cause false positive results in the differential analysis. In other words, we remove genes that only express in a cell type but does not express in the majority of cells in the underlying cell type.

Next, we apply the Trimmed Mean of M-values (TMM) normalization to the filtered data [20]. After the filtering and normalization steps, we perform differential analysis to identify the markers for each known cell type. For a specific cell type, we use the empirical Bayes statistics implemented in limma [21] to identify the differentially expressed genes (DE genes) by comparing the expression of cells in one cell type against all of the remaining cells in other cell types. The threshold for DE gene is having an adjusted p-value smaller than 0.05 (using Benjamini-Hochberg's). These DE genes are then used as the markers for that cell type. We repeat this process to obtain a list of markers for all cell types.

After obtaining the list of markers for each cell type, we compute the signature matrix. For a specific cell type and a marker, the expression of the cell type is calculated as the average expression of the cells belonging to the cell type. Repeating this procedure for all cell types and all markers, we obtain a signature matrix of markers by cell types. This matrix serves as the input of the next module to estimate the cell type proportion in each bulk sample.

B. Module II: Cell type proportions estimation

After obtaining the cell type signature matrix, we estimate the proportions of cell type in each sample in bulk expression data by utilizing a novel application of elastic net regression. First, we normalize the bulk expression data using DESeq2 [22]. In particularly, we use DESeq2 to estimate the size factors of each sample using the median ratio method, and then normalize the bulk data [23]. We also perform an integrity check such that the bulk data and the obtained signature matrix contain the same set of genes. Next, for each bulk sample, we model the gene expression as a linear combination of its cell type-specific expressions, in which the coefficients are the proportions of the cell types presenting in the sample.

To estimate these coefficients, we build a linear regression model to predict the bulk gene expression from the cell type signature profiles. Here we propose *weighted elastic net*, which is an extension of *elastic net* model [24]. *Elastic net* is a well-known regularized regression technique that automatically selects the features (or cell types) for building the linear regression model, by imposing L1-norm and L2-norm penalties on the regression coefficients. Here we extend this framework by specifying weight for each gene in calculating loss function in the objective function of *elastic net*. In our model, we set the weight of gene i as $w_i = \frac{1}{\sum_{j=1}^{N} S_{ij}+1}$, where S is the signature matrix, N is the number of cell types. These weights help prevent the model from only focusing on reducing the residual errors in the highly expressed genes and ignoring the rest, thereby improve model's robustness.

Taken together, the *weighted elastic net* proposed in DWEN has addressed many advantages over current cellular deconvolution methods: (i) robustness to noise as the weights prevent the model from only focusing on reducing the residual errors in the highly expressed genes and ignoring the rest; (ii) automatic feature selection from the signature matrix using L1-norm function; and (iii) better prediction performance as tolerance to multi-collinearity via utilization of the L2-norm penalty function. Additionally, because the proportions are positive, we also apply a non-negative constraint when estimating the model's coefficients. Finally, the coefficients in the model are used as the proportions of the corresponding cell types in the bulk sample. For this purpose, we apply the coordinate descent algorithm [25] to optimally solve our objective function. Specifically, the algorithm starts with initializing random values for the model's coefficients. It then alternatively updates one coefficient via setting the gradient of objective function with respect to this coefficient to 0, while keeping the others fixed. The algorithm iteratively updates and calculates these coefficients until it reaches convergence.

III. Result

To evaluate the accuracy of the deconvolution methods, we perform a comprehensive simulation study using single-cell RNA sequencing data from the human cell atlas, Tabula Sapiens [19]. The processed data are downloaded from the atlas website (https://tabula-sapiens-portal.ds.czbiohub.org/).

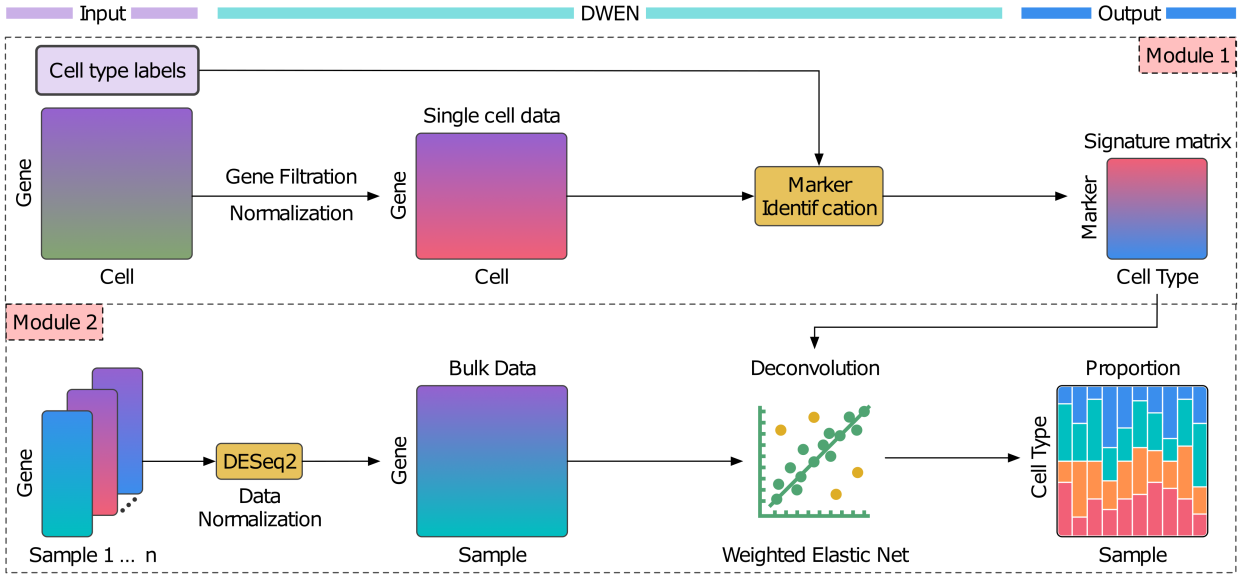

Fig. 1. The overall workflow of Deconvolution using Weighted Elastic Net (DWEN). The input of the method includes: (i) a reference single-cell dataset of known cell types, and (ii) a bulk expression dataset to be deconvolved. The method consists of two main modules: (i) signature matrix construction, and (ii) cell type proportions estimation. The first module first performs gene filtering and normalization to reduce noise and then performs differential analysis to identify the important markers for each cell type. It then aggregates the expression of cells belonging to each cell type to obtain the signature matrix (rows are markers and columns are known cell types). The second module aims to estimate the proportion of each cell type in the bulk sample using weighted elastic net. It uses DESeq2 to estimate the size factors of each samples and then normalizes the bulk data. For each bulk sample, the model trains a weighted elastic net with cell type signature profiles as the predictors, and the bulk expression profiles as the outcome. The weights of the predictors are considered as the proportions of the corresponding cell types in the bulk sample.

This dataset consists of approximately 500,000 cells from 45 tissues of 15 normal human subjects. We select 20 tissues that have the data from at least two donors for our benchmark. The details of the data are provided in Table I.

For a specific tissue, we choose the scRNA-seq data from one donor as the reference single-cell dataset and use the scRNA-seq data from the remaining donors to simulate bulk data. To simulate a bulk sample, we first generate random cell type proportions, which sum up to one. We set the number of cells in each bulk sample to be 5,000 cells. We then randomly select the cells from scRNA-seq data to match the defined proportions. For example, if a cell type has a proportion of 0.1, we select 500 cells of that cell type. The expression of the bulk sample is the sum of the expression of all 5,000 cells. We repeat the same process to generate 100 bulk samples for the tissue. Repeating the whole procedure for each of the 20 tissues, we obtain 20 datasets – one per tissue. In these bulk datasets, the true cell type proportions are known and thus can be used to assess the performance of deconvolution methods.

For each tissue/dataset, we use the six deconvolution methods to infer the cell type proportions: DWEN, EPIC [16], SCDC [18], DAISM [15], TOAST [13], and BayesCCE [10]. We use the reference single-cell expression matrix to construct the cell type signature matrix as described in the Method section. We also use the signature matrix generated by DWEN for the input of EPIC and SCDC because these two methods do not support the signature matrix construction. Note that BayesCCE and TOAST are reference-free methods while DWEN, EPIC, SCDC, and DAISM are reference-based approaches. To quantify the accuracy of each method, we calculate the Spearman correlation between the estimated cell type proportions and the true proportions. A good deconvolution method is expected to have a high correlation in each dataset.

Figure 2 shows the analysis results of the Eye dataset. The most top left panel shows the results of DWEN. Each data point represents a cell type in a sample. For example, a red point represents the "conjunctival_epithelial_cell" type in a sample while points of other colors represent other cell types. The horizontal axis shows the true proportion (from ground truth) of the cell type while the vertical axis shows the estimated proportion. The panel shows that DWEN accurately estimates the proportion of most cell types in most samples. The correlation between the estimated proportion and true proportion is 0.81. The other five methods, EPIC, SCDC, DAISM, TOAST, and BayesCCE, are substantially less accurate than DWEN. Their correlations are 0.62, 0.47, 0.55, 0.03, and 0.55, respectively. Notable EPIC performs the best among the five existing methods but it lacks consistency. It performs well in the majority of samples but fails to estimate the proportions in the remaining samples. Nevertheless, DWEN outperforms all current methods in this dataset.

Table II shows the Spearman correlation of all six methods in 20 datasets. Cells highlighted in bold text have the highest correlation in the corresponding row. Overall, DWEN outperforms other methods by achieving the highest correlation in 16/20 datasets (all except Fat, Thymus, Small Intestine, and Bone Marrow). It also has a substantially higher average correlation than other competing methods.

TABLE I
DESCRIPTION OF THE 20 TISSUES INCLUDED IN THE BENCHMARK.

Tissue	#Donor	#Cell Type	#Genes	#UMI	Cell Types
Bladder	3	9	2,739	13,219	T cell, macrophage, myofibroblast cell, bladder urothelial cell, smooth muscle cell, fibroblast, pericyte cell, mast cell, mature NK T cell
Blood	6	6	1,866	9,100	erythrocyte, classical monocyte, neutrophil, memory B cell, plasma cell, platelet
Bone Marrow	3	8	2,600	11,848	plasma cell, hematopoietic stem cell, erythroid progenitor cell, mature NK T cell, granulocyte, naive B cell, CD8 positive alpha beta T cell, CD4 positive alpha beta T cell
Eye	3	7	3,286	17,357	conjunctival epithelial cell, eye photoreceptor cell, Muller cell, retinal blood vessel endothelial cell, keratocyte, corneal epithelial cell, melanocyte
Fat	2	4	3,247	13,353	fibroblast, endothelial cell, macrophage, myofibroblast cell
Large Intestine	2	5	3,764	16,385	CD8 positive alpha beta T cell, fibroblast, paneth cell of colon, B cell, gut endothelial cell
Liver	2	2	2,729	10,123	endothelial cell of hepatic sinusoid, hepatocyte
Lung	3	3	1,849	9,102	type II pneumocyte, mature NK T cell, adventitial cell
Lymph Node	3	9	2,302	8,458	B cell, effector CD4 positive alpha beta T cell, regulatory T cell, plasma cell, neutrophil, macrophage, CD1c positive myeloid dendritic cell, intermediate monocyte, mast cell
Muscle	3	11	3,282	15,256	mesenchymal stem cell, skeletal muscle satellite stem cell, capillary endothelial cell, pericyte cell, fast muscle cell, macrophage, endothelial cell of vascular tree, slow muscle cell, endothelial cell of artery, tendon cell, endothelial cell of lymphatic vessel
Pancreas	2	7	2,024	7,477	pancreatic acinar cell, T cell, endothelial cell, myeloid cell, pancreatic stellate cell, B cell, pancreatic ductal cell
Prostate	2	6	2,532	10,319	basal cell of prostate epithelium, epithelial cell, club cell, erythroid progenitor cell, luminal cell of prostate epithelium, endothelial cell
Salivary Gland	2	10	2,564	9,155	acinar cell of salivary gland, pericyte cell, mature NK T cell, fibroblast, endothelial cell of lymphatic vessel, adventitial cell, endothelial cell, monocyte, duct epithelial cell, basal cell
Skin	2	8	3,031	19,725	macrophage, stromal cell, mast cell, muscle cell, CD1c positive myeloid dendritic cell, endothelial cell, naive thymus derived CD8 positive alpha beta T cell, regulatory T cell
Small Intestine	2	4	2,480	10,034	CD8 positive alpha beta T cell, B cell, paneth cell of epithelium of small intestine, fibroblast
Spleen	3	13	2,475	13,680	macrophage, intermediate monocyte, endothelial cell, memory B cell, classical monocyte, neutrophil, naive B cell, plasma cell, type I NK T cell, mature NK T cell, innate lymphoid cell, regulatory T cell, hematopoietic stem cell
Thymus	2	9	2,160	8,746	medullary thymic epithelial cell, fibroblast, macrophage, vascular associated smooth muscle cell, plasma cell, vein endothelial cell, capillary endothelial cell, endothelial cell of artery, monocyte
Tongue	2	5	1,971	8,706	leukocyte, fibroblast, vein endothelial cell, pericyte cell, capillary endothelial cell
Trachea	2	3	2,395	9,850	endothelial cell, ciliated cell, basal cell
Vasculature	2	6	2,414	8,794	fibroblast, smooth muscle cell, macrophage, pericyte cell, mast cell, mature NK T cell

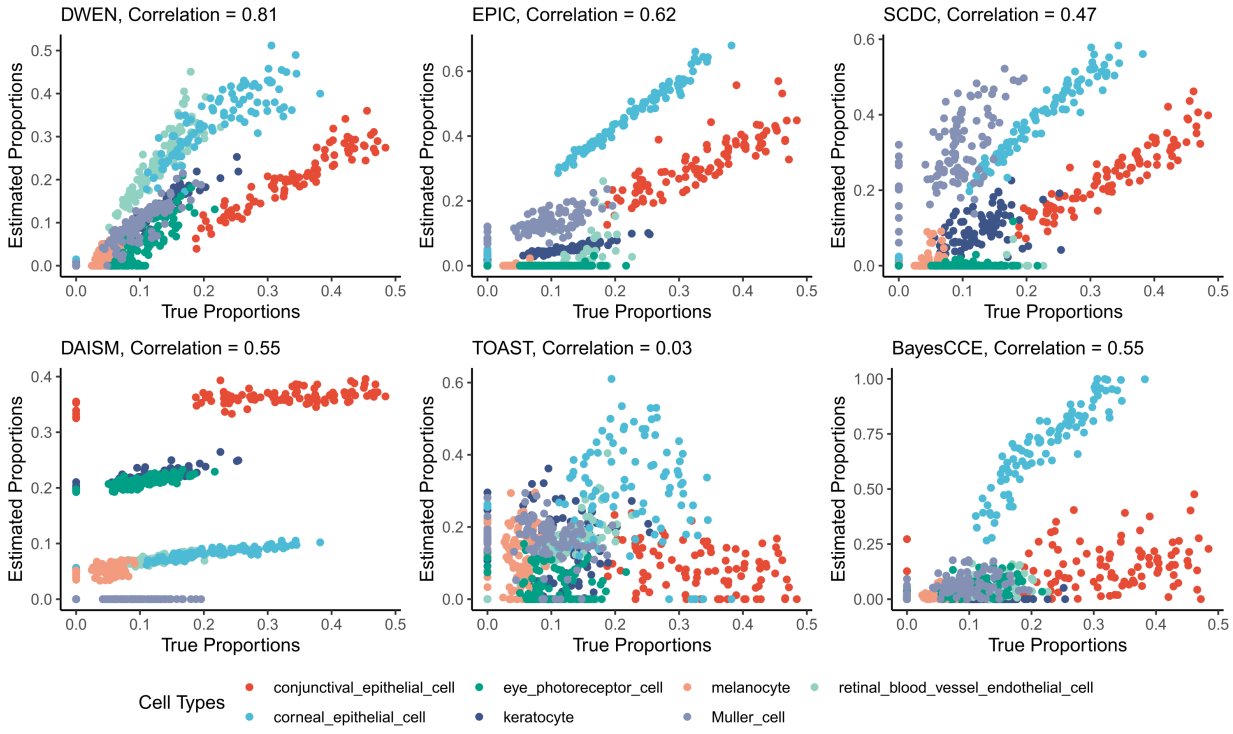

Fig. 2. Evaluation of deconvolution methods using the Eye dataset. The horizontal axis shows the true cell type proportions while the vertical axis shows the cell type proportions estimated by DWEN, EPIC, SCDC, DAISM, TOAST, and BayesECE. A point on a panel represents a cell type in a sample (there are a total of 100 samples in this dataset). The performance of a method is quantified by the correlation between the estimated cell type proportions and the true proportions. DWEN outperforms all state-of-the-at methods by having the highest correlation.

TABLE II
ACCURACY OF CELL TYPE PROPORTIONS INFERRED BY BAYESCCE, TOAST, DAISM, EPIC, SCDC, AND DWEN MEASURED BY SPEARMAN CORRELATION. THE HIGHEST VALUES ARE HIGHLIGHTED IN BOLD.

Tissue	DWEN	EPIC	SCDC	DAISM	TOAST	BayesCCE
Liver	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	-1.00
Trachea	**0.90**	0.76	-0.02	0.78	0.24	0.05
Lung	**0.90**	0.60	0.86	0.86	0.80	0.63
Blood	**0.89**	0.49	**0.89**	0.60	0.74	0.35
Vasculature	**0.84**	0.82	0.50	0.12	0.42	-0.11
Lymph Node	**0.83**	0.49	0.37	0.24	0.23	0.20
Bladder	**0.82**	0.52	0.49	0.08	-0.06	0.15
Eye	**0.81**	0.62	0.47	0.55	0.03	0.55
Spleen	**0.80**	0.69	0.70	0.47	0.02	0.31
Muscle	**0.79**	0.63	0.57	0.29	0.31	0.42
Skin	**0.77**	0.38	-0.07	0.50	0.42	0.09
Tongue	**0.77**	0.59	0.71	0.63	0.70	-0.13
Pancreas	**0.76**	0.58	0.62	0.24	-0.07	0.18
Large Intestine	**0.73**	0.62	0.58	0.12	0.02	0.17
Fat	0.61	0.48	0.03	**0.79**	0.41	0.47
Salivary Gland	**0.60**	**0.60**	0.44	0.49	0.41	0.19
Thymus	0.54	0.20	0.35	0.07	**0.61**	0.26
Small Intestine	0.53	-0.41	0.04	**0.66**	0.24	0.21
Prostate	**0.52**	0.42	0.48	-0.39	0.23	-0.09
Bone Marrow	0.41	0.14	0.26	0.17	**0.44**	0.03
Mean	**0.74**	0.51	0.46	0.41	0.36	0.15

For the Liver dataset, most methods except BayesCCE can accurately estimate the cell type proportions with a perfect correlation of 1. For the Trachea dataset, the three methods DWEN, EPIC, and DAISM perform well with correlations of 0.75 and above. The other three methods, SCDC, TOAST, and BayesCCE, have correlations that are close to zero. In other words, these methods are not ideal for this specific tissue. For the Lung dataset, most methods perform well with DWEN being the best method.

The same trend can be observed in the remaining datasets: DWEN consistently outperforms other methods in most analyses. DWEN substantially outperforms all other methods by having the highest average correlation. The average correlation of DWEN across all 20 datasets is 0.74 while that of BayesCCE, TOAST, DAISM, EPIC, and SCDC are 0.15, 0.36, 0.41, 0.51, and 0.46, respectively. DWEN also has the highest correlations in most datasets. This demonstrates that DWEN can accurately and reliably estimate the proportions of the cell types in all samples and all tissues.

IV. CONCLUSION

In this article, we introduced a new method, DWEN, to infer the cell type proportions of bulk expression data using single-cell expression data as reference. We compared DWEN with five state-of-the-art deconvolution methods using 20 datasets obtained from the human cell atlas. We demonstrated that DWEN outperforms other methods in inferring the cell type proportions of the bulk samples. A potential improvement of this research is to develop an ensemble deconvolution approach when multiple signature matrices are available. We also plan to combine DWEN with our current analysis techniques in these applications, including cancer subtyping [26–33], meta-analysis [34–37], single-cell analysis [38–41], and systems-level interpretation [42–48], and the analysis of omics data other than transcriptome [49–51].

V. ACKNOWLEDGMENTS

This work was partially supported by NSF under grant numbers 2141660, 2203236, 2001385, and 2019609. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

REFERENCES

[1] D. S. Chen and I. Mellman. Elements of cancer immunity and the cancer–immune set point. *Nature*, 541:321–330, 2017.

[2] M. R. Junttila and F. J. de Sauvage. Influence of tumour microenvironment heterogeneity on therapeutic response. *Nature*, 501:346–354, 2013.

[3] A. J. Gentles, A. M. Newman, C. L. Liu, S. V. Bratman, W. Feng, D. Kim, V. S. Nair, Y. Xu, A. Khuong, C. D. Hoang, M. Diehn, R. B. West, S. K. Plevritis, and A. A. Alizadeh. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nature Medicine*, 21:938–945, 2015.

[4] O. Cabral-Marques, L. F. Schimke, E. B. de Oliveira Jr., N. El Khawanky, R. N. Ramos, B. K. Al-Ramadi, G. R. S. Segundo, H. D. Ochs, and A. Condino-Neto. Flow cytometry contributions for the diagnosis and immunopathological characterization of primary immunodeficiency diseases with immune dysregulation. *Frontiers in Immunology*, 10:2742, 2019.

[5] L. C. Lawrie, S. Curran, H. L. McLeod, J. E. Fothergill, and G. I. Murray. Application of laser capture microdissection and proteomics in colon cancer. *Molecular Pathology*, 54(4):253–258, 2001.

[6] J. Duraiyan, R. Govindarajan, K. Kaliyappan, and M. Palanisamy. Applications of immunohistochemistry. *Journal of pharmacy & bioallied sciences*, 4(Suppl 2):S307, 2012.

[7] R. Sandberg. Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods*, 11:22–24, 2014.

[8] G. Li, Y. Jiang, G. Li, and Q. Qiao. Comprehensive analysis of radiosensitivity in head and neck squamous cell carcinoma. *Radiotherapy and Oncology*, 159:126–135, 2021.

[9] F. Xie, M. Zhou, and Y. Xu. BayCount: A Bayesian decomposition method for inferring tumor heterogeneity using RNA-Seq counts. *The Annals of Applied Statistics*, 12(3):1605–1627, 2018.

[10] E. Rahmani, R. Schweiger, L. Shenhav, T. Wingert, I. Hofer, E. Gabel, E. Eskin, and E. Halperin. BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference. *Genome Biology*, 19:141, 2018.

[11] L. A. Newberg, X. Chen, C. D. Kodira, and M. I. Zavodszky. Computational de novo discovery of distinguishing genes for biological processes and cell types in complex tissues. *PLoS ONE*, 13(3):e0193067, 2018.

[12] D. Repsilber, S. Kern, A. Telaar, G. Walzl, G. F. Black, J. Selbig, S. K. Parida, S. H. E. Kaufmann, and M. Jacobsen. Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach. *BMC Bioinformatics*, 11:27, 2010.

[13] Z. Li and H. Wu. TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome Biology*, 20:190, 2019.

[14] T. Yang, N. Alessandri-Haber, W. Fury, M. Schaner, R. Breese, M. LaCroix-Fralish, J. Kim, C. Adler, L. E. Macdonald, G. S. Atwal, and Y. Bai. AdRoit is an accurate and robust method to infer complex transcriptome composition. *Communications Biology*, 4:1218, 2021.

[15] Y. Lin, H. Li, X. Xiao, L. Zhang, K. Wang, J. Zhao, M. Wang, F. Zheng, M. Zhang, W. Yang, J. Han, and R. Yu. DAISM-DNNXMBD: Highly accurate cell type proportion estimation

with in silico data augmentation and deep neural networks. *Patterns*, 3(3):100440, 2022.

[16] J. Racle, K. de Jonge, P. Baumgaertner, D. E. Speiser, and D. Gfeller. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife*, 6:e26476, 2017.

[17] F. Finotello, C. Mayer, C. Plattner, G. Laschober, D. Rieder, H. Hackl, A. Krogsdam, Z. Loncova, W. Posch, D. Wilflingseder, S. Sopper, M. Ijsselsteijn, T. P. Brouwer, D. Johnson, Y. Xu, Y. Wang, M. E. Sanders, M. V. Estrada, P. Ericsson-Gonzalez, P. Charoentong, J. Balko, N. F. d. C. C. de Miranda, and Z. Trajanoski. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Medicine*, 11:34, 2019.

[18] M. Dong, A. Thennavan, E. Urrutia, Y. Li, C. M. Perou, F. Zou, and Y. Jiang. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Briefings in Bioinformatics*, 22(1):416–427, 2021.

[19] T. S. Consortium. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022.

[20] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[21] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 01 2015.

[22] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550, 2014.

[23] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[24] H. Zou and T. Hastie. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[25] S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151:3–34, 2015.

[26] T. Nguyen, R. Tagett, D. Diaz, and S. Draghici. A novel approach for data integration and disease subtyping. *Genome Research*, 27:2025–2039, 2017.

[27] H. Nguyen, S. Shrestha, S. Draghici, and T. Nguyen. PINSPlus: A tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*, 35(16):2843–2846, 2019.

[28] D. Tran, H. Nguyen, U. Le, G. Bebis, H. N. Luu, and T. Nguyen. A novel method for cancer subtyping and risk prediction using consensus factor analysis. *Frontiers in Oncology*, 10:1052, 2020.

[29] H. Nguyen, D. Tran, B. Tran, M. Roy, A. Cassell, S. Dascalu, S. Draghici, and T. Nguyen. SMRT: Randomized data transformation for cancer subtyping and big data analysis. *Frontiers in Oncology*, 11:725133, 2021.

[30] T. H. Y. Nguyen, T. Nguyen, Q.-H. Nguyen, and D.-H. Le. Re-identification of patient subgroups in uveal melanoma. *Frontiers in Oncology*, 11:731548, 2021.

[31] Q.-H. Nguyen, H. Nguyen, T. Nguyen, and D.-H. Le. Multi-omics analysis detects novel prognostic subgroups of breast cancer. *Frontiers in Genetics*, 11:1265, 2020.

[32] Q.-H. Nguyen, T. Nguyen, and D.-H. Le. Identification and validation of a novel three hub long noncoding RNAs with m6A modification signature in low-grade gliomas. *Frontiers in Molecular Biosciences*, 9:801931, 2022.

[33] Q.-H. Nguyen, T. Nguyen, and D.-H. Le. DrGA: cancer driver gene analysis in a simpler manner. *BMC Bioinformatics*, 23:86, 2022.

[34] T. Nguyen, R. Tagett, M. Donato, C. Mitrea, and S. Draghici. A novel bi-level meta-analysis approach-applied to biological pathway analysis. *Bioinformatics*, 32(3):409–416, 2016.

[35] T. Nguyen, C. Mitrea, R. Tagett, and S. Draghici. DANUBE: Data-driven meta-ANalysis using UnBiased Empirical distributions - applied to biological pathway analysis. *Proceedings of the IEEE*, 105(3):496–515, 2017.

[36] T. Nguyen, D. Diaz, R. Tagett, and S. Draghici. Overcoming the matched-sample bottleneck: an orthogonal approach to integrate omic data. *Scientific Reports*, 6:29251, 2016.

[37] T. Nguyen, A. Shafi, T.-M. Nguyen, A. G. Schissler, and S. Draghici. NBIA: a network-based integrative analysis framework–applied to pathway analysis. *Scientific Reports*, 10:4188, 2020.

[38] D. Tran, H. Nguyen, B. Tran, C. La Vecchia, H. N. Luu, and T. Nguyen. Fast and precise single-cell data analysis using hierarchical autoencoder. *Nature Communications*, 12:1029, 2021.

[39] D. Tran, B. Tran, H. Nguyen, and T. Nguyen. A novel method for single-cell data imputation using subspace regression. *Scientific Reports*, 12:2697, 2022.

[40] B. Tran, D. Tran, H. Nguyen, S. Ro, and T. Nguyen. scCAN: single-cell clustering using autoencoder and network fusion. *Scientific Reports*, 12:10267, 2022.

[41] H. Nguyen, D. Tran, B. Tran, B. Pehlivan, and T. Nguyen. A comprehensive survey of regulatory network inference methods using single-cell RNA sequencing data. *Briefings in Bioinformatics*, 22(3):1–15, 2021.

[42] H. Nguyen, D. Tran, J. M. Galazka, S. V. Costes, A. Beheshti, S. Draghici, and T. Nguyen. CPA: A web-based platform for consensus pathway analysis and interactive visualization. *Nucleic Acids Research*, 49(W1):W114–W124, 2021.

[43] T.-M. Nguyen, A. Shafi, T. Nguyen, and S. Draghici. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biology*, 20:203, 2019.

[44] H. Nguyen, S. Shrestha, D. Tran, A. Shafi, S. Draghici, and T. Nguyen. A comprehensive survey of tools and software for active subnetwork identification. *Frontiers in Genetics*, 10:155, 2019.

[45] A. Shafi, T. Nguyen, A. Peyvandipour, H. Nguyen, and S. Draghici. A multi-cohort and multi-omics meta-analysis framework to identify network-based gene signatures. *Frontiers in Genetics*, 10:159, 2019.

[46] A. Shafi, T. Nguyen, A. Peyvandipour, and S. Draghici. GSMA: an approach to identify robust global and test Gene Signatures using Meta-Analysis. *Bioinformatics*, 36(2):487–495, 2019.

[47] E. Cruz, H. Nguyen, T. Nguyen, and I. Wallace. Functional analysis tools for post-translational modification: a post-translational modification database for analysis of proteins and metabolic pathways. *The Plant Journal*, 99(5):1003–1013, 2019.

[48] T. Nguyen, C. Mitrea, and S. Draghici. Network-based approaches for pathway level analysis. *Current Protocols in Bioinformatics*, 61(1):8–25, 2018.

[49] B. T. Caswell, C. C. de Carvalho, H. Nguyen, M. Roy, T. Nguyen, and D. C. Cantu. Thioesterase enzyme families: Functions, structures, and mechanisms. *Protein Science*, 31(3):652–676, 2022.

[50] J. C. Stansfield, D. Tran, T. Nguyen, and M. G. Dozmorov. R tutorial: Detection of differentially interacting chromatin regions from multiple Hi-C datasets. *Current Protocols in Bioinformatics*, 66(1):e76–e76, 2019.

[51] A. Shafi, C. Mitrea, T. Nguyen, and S. Draghici. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Briefings in Bioinformatics*, 19(5):737–753, 2018.