JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 27, Number 3, 2020
Mary Ann Liebert, Inc.
Pp. 386–389
DOI: 10.1089/cmb.2019.0469

DTD: An R Package for Digital Tissue Deconvolution

MARIAN SCHÖN,[1] JAKOB SIMETH,[1] PAUL HEINRICH,[1] FRANZISKA GÖRTLER,[1]
STEFAN SOLBRIG,[2] TILO WETTIG,[2] PETER J. OEFNER,[3]
MICHAEL ALTENBUCHINGER,[1] and RAINER SPANG[1]

ABSTRACT

Digital tissue deconvolution (DTD) estimates the cellular composition of a tissue from its bulk gene-expression profile. For this, DTD approximates the bulk as a mixture of cell-specific expression profiles. Different tissues have different cellular compositions, with cells in different activation states, and embedded in different environments. Consequently, DTD can profit from tailoring the deconvolution model to a specific tissue context.

Loss-function learning adapts DTD to a specific tissue context, such as the deconvolution of blood, or a specific type of tumor tissue. We provide software for loss-function learning, for its validation and visualization, and for applying the DTD models to new data.

Keywords: cell-type deconvolution, loss-function learning, model adaptation, R package.

1. INTRODUCTION

The cellular composition of a tumor specimen is a prognostic factor (Ansell and Vonderheide, 2013; Junttila and de Sauvage, 2013). Single-cell RNA sequencing (scRNA-Seq; Wu et al., 2013) can be used to assess this composition experimentally. Digital tissue deconvolution (DTD) emerged as a computational alternative (Cobos et al., 2018). It can be applied retrospectively to bulk gene-expression data without experimental costs.

Let y be a bulk gene expression profile, X a matrix with cell-type specific reference profiles in its columns, and c a vector of cellular proportions. DTD reconstructs c through X and y by $y = Xc$, where different objective functions can be used to approximate c (Mohammadi et al., 2017). One of them is the sum of squared residuals, $||y - Xc||_2^2$, where the residuals are calculated between observed gene expression y and its reconstruction Xc.

An important observation is that genes can be weighted to obtain better estimates (Görtler et al., 2018). This is particularly the case (1) if not all cells in the bulk are represented in X, (2) if we want to estimate contributions from small cell populations, and (3) if we want to disentangle highly similar cell types. Genes weights can be introduced by replacing the residual sum of squares by

[1]Department of Statistical Bioinformatics, Institute of Functional Genomics, University of Regensburg, Regensburg, Germany.
[2]Department of Physics, University of Regensburg, Regensburg, Germany.
[3]Institute of Functional Genomics, University of Regensburg, Regensburg, Germany.

$$\arg\min_{c} ||\mathrm{diag}(g)(y - Xc)||_2^2, \tag{1}$$

with a vector g of gene weights $g_i \geq 0$. A high weight g_i for gene i indicates that it is important for deconvolution, whereas a low weight corresponds to a gene that deteriorates the deconvolution. For example, if the expression of a gene in the tissue differs greatly from the corresponding expression in the reference profiles, it cannot be explained by a linear deconvolution approach.

Loss-function learning (Görtler et al., 2018) detects this problem and adapts g to a tissue context. It increases deconvolution accuracy, as shown exemplarily for the deconvolution of bulk melanoma specimens (Görtler et al., 2018).

2. METHODS

Loss-function learning uses training data to optimize DTD models. These training data are bulk gene-expression measurements Y (rows: genes/features, columns: measurements) with known cellular compositions C (rows: cellular proportions, columns: measurements). The rational behind loss-function learning is to obtain the vector g by minimizing a loss function L on the training data.

The package DTD incorporates a correlation-based loss function:

$$\begin{aligned} L &= -\sum_{j=1}^{q} \mathrm{cor}\left(C_{j,.}, \hat{C}_{j,.}(g)\right), \\ &\text{subject to } g_i \geq 0 \text{ and} ||g||_2 = 1 \end{aligned} \tag{2}$$

Here, $\hat{C}(g)$ is the solution of Equation (1). The optimization problem is stated in a closed analytical form and is implemented efficiently in C++ for optimal performance.

3. APPLICATION

3.1. Basic function calls

The R package DTD has the basic function call

$$\text{train\_deconvolution\_model(X.matrix = X, train.data.list = train.data, } \ldots) \tag{3}$$

Here, X is a matrix with reference profiles in its columns, and train.data is a list containing ''mixtures'' Y and ''quantities'' C, defined as earlier. train_deconvolution_model returns a deconvolution model that can be applied through the function estimate_c(DTD.model = trained.model, new.data = y) on new data y, which is a vector or matrix with bulk gene-expression levels. The function estimate_c returns estimated cellular proportions for the bulk profiles y. The workflow is summarized in Figure 1.

3.2. Generate artificial training data from scRNA-Seq data

Loss-function learning requires training data. These data can be generated experimentally, for example, through fluorescence-activated cell sorting (FACS)-based cell counting combined with bulk RNA sequencing. Alternatively, these data can be generated artificially from scRNA-Seq experiments. For this purpose, we implemented functions that automatically generate training mixtures Y and their corresponding compositions C. An example for such a function call is mix_samples(exp.data = sc.counts, pheno = sc.pheno, …). This function uses, for example, raw counts from scRNA-Seq (sc.counts), where columns correspond to cells and rows to genes. Furthermore, the vector sc.pheno ascribes cell types to the respective columns in sc.counts. The function returns a list with the components Y and C, which correspond to artificial mixtures and their underlying cellular compositions, respectively. A similar function is provided to generate a reference matrix X.

4. SUMMARY

We provide software for the digital deconvolution of bulk gene-expression data. We take into account that DTD needs to be adapted to a specific type of tissue. It is not likely that there is a universal

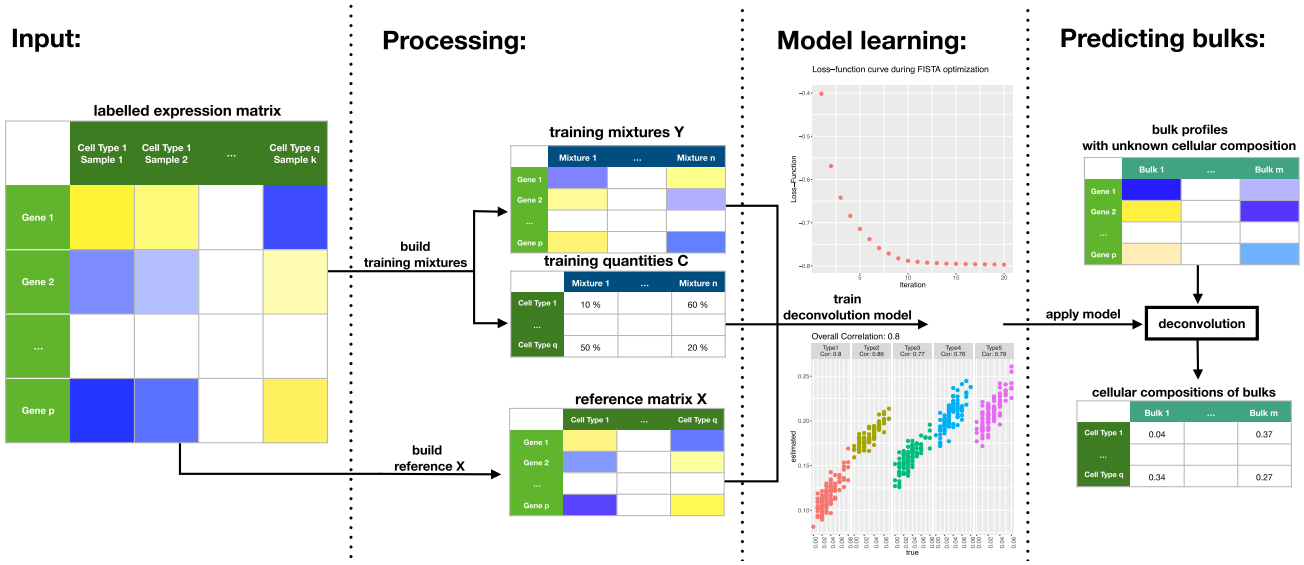

FIG. 1. Workflow of a digital tissue deconvolution (DTD) analysis with loss-function learning. Input: an expression matrix, where each sample is labeled with its cell type. Processing: the labeled samples are used to build both a reference matrix and artificial mixtures of known cellular composition. Model learning: the algorithm iteratively searches for parameters g, which maximize the correlation between the estimated and the true cellular compositions, where the training data and the reference matrix are used. Here, functions visualize the result and assess the performance of the DTD model. Predicting bulks: the DTD model is applied to bulk gene expression data to estimate the underlying cellular composition.

deconvolution formula, which performs consistently well in all settings. The most reliable results might be obtained by optimized models that have been trained within a machine learning framework.

In this study, we provide software for such a framework, called loss-function learning, and for the application of the learned deconvolution models to bulk data. It is computationally tractable and easy to use.

A DTD tutorial is available as Supplementary Data. There, we give a comprehensive example that shows how data need to be processed and how results can be visualized.

In summary, we provide the R package DTD that contains software to systematically improve the performance of DTD algorithms.

AVAILABILITY AND IMPLEMENTATION

DTD is available under https://github.com/MarianSchoen/DTD.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

This study was supported by BMBF (TissueResolver 031L0173) and DFG (FOR-2127 and SFB/TRR-55).

SUPPLEMENTARY MATERIAL

Supplementary Data

REFERENCES

Ansell, S.M., and Vonderheide, R.H. 2013. Cellular composition of the tumor microenvironment. *Am. Soc. Clin. Oncol. Educ. Book* 33, e91–e97.

Cobos, F.A., Vandesompele, J., Mestdagh, P., et al. 2018. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* 34, 1969–1979.

Görtler, F., Solbrig, S., Wettig, T., et al. 2018. *Research in Computational Molecular Biology: 22nd Annual International Conference, RECOMB 2018, Paris, France, April 21–24, 2018, Proceedings (Lecture Notes in Computer Science)*. Springer.

Görtler, F., Schön, M., Simeth, J., et al. 2020. Loss-function learning for digital tissue deconvolution. *J. Comp. Biol*. 27, this issue.

Junttila, M.R., and de Sauvage, F.J. 2013. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature* 501, 346–354.

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

Wu, A.R., Neff, N.F., Kalisky, T., et al. 2013. Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* 11, 41–46.

Address correspondence to:
Marian Schön
Statistical Bioinformatics
Institute of Functional Genomics
University of Regensburg
Am BioPark 9
Regensburg 93053
Germany

E-mail: marian.schoen@klinik.uni-regensburg.de