Adaptive Digital Tissue Deconvolution

Franziska Görtler[1], Malte Mensching-Buhr[2], Ørjan Skaar[3], Stefan Schrod[2], Tim Beißbarth[2], Anagha Joshi[4], Helena U. Zacharias[5], Sushma Nagaraja Grellscheid[1], and Michael Altenbuchinger[2]

[1]Computational Biology Unit, Department of Biological Sciences, University of Bergen, Thormohlensgt 55, Bergen N-5008, Norway.
[2]Department of Medical Bioinformatics, University Medical Center Göttingen, 37075 Göttingen, Germany.
[3]Department of Informatics, Computational Biology Unit, University of Bergen, Thormohlensgt 55, Bergen N-5008, Norway.
[4]Department of Clinical Science, Computational Biology Unit, University of Bergen, Thormohlensgt 55, Bergen N-5008, Norway.
[5]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, 30625 Hannover, Germany.

correspondence to: michael.altenbuchinger@bioinf.med.uni-goettingen.de and sushma.grellscheid@uib.no

Abstract

Motivation: The inference of cellular compositions from bulk and spatial transcriptomics data increasingly complements data analyses. Multiple computational approaches were suggested and recently, machine learning techniques were developed to systematically improve estimates. Such approaches allow to infer additional, less abundant cell types. However, they rely on training data which do not capture the full biological diversity encountered in transcriptomics analyses; data can contain cellular contributions not seen in the training data and as such, analyses can be biased or blurred. Thus, computational approaches have to deal with unknown, hidden contributions. Moreover, most methods are based on cellular prototypes which serve as a reference; e.g., a generic T-cell profile is used to infer the proportion of T-cells. It is well known that cells adapt their molecular phenotype to the environment and as such, pre-specified cell prototypes can distort the inference of cellular compositions.
Results: We propose Adaptive Digital Tissue Deconvolution (ADTD) to estimate cellular proportions of pre-selected cell types together with possibly unknown and hidden background contributions. Moreover, ADTD adapts the prototypic reference profiles to the molecular environment of the cells, which allows one to resolve cell-type specific regulation from bulk transcriptomics data. The performance of ADTD was verified in simulation studies, and in an application to breast cancer data we demonstrate how ADTD can be used to gain insights into molecular differences between breast cancer sub-types.
Availability and implementation: A python implementation of ADTD and a tutorial

are available at zenodo (doi:10.5281/zenodo.7548362).
Contact:michael.altenbuchinger@bioinf.med.uni-goettingen.de
Supplementary information: Supplementary material is available at *Bioinformatics* online.

1 Introduction

Bulk transcriptomics profiles are a complex linear combination of diverse molecular contributions of multiple individual cells. Many analyses such as the screening for differentially expressed genes can be confounded by the underlying cellular compositions and consequently, if the cellular proportions are unknown, we cannot statistically resolve the source of differential expression. Multiple approaches were suggested to estimate cellular proportions from bulk transcriptomics data, ranging from traditional, regression-based approaches to machine-learning techniques, including generalized regression [1, 2, 3] as well as deep learning approaches [4, 5]. In recent years, the advent of spatial transcriptomics has generated an additional need for reliable cell-type deconvolution [6, 7].

Single-cell RNA sequencing data can substantially improve estimates of cellular proportions via improved reference matrices as well as collections thereof, and via optimized gene selection or weighting, as demonstrated by work of us and others [8, 9, 10, 11, 12]. The latter is usually facilitated by generating artificial mixtures of known cellular composition which are used to assess the model performance by comparing the ground truth cellular composition with its estimate. Using such a strategy, for instance, MuSiC [9] and DTD [10] learn a gene weighting and Bisque [11] learns gene-specific bulk expression transformations.

A recent comprehensive benchmarking of different methods for cell-type deconvolution resolved the marker gene selection as a major determinant of model performance and it was pointed out that cell types missing in the reference profiles substantially worsen results [13]. The latter statement is in line with the observation that MuSiC, Non-negative Least Squares (NNLS), and CIBERSORT [1] did not produce accurate results in a validation study using artificial bulk data of six pancreatic cell types if one cell type is removed from the reference single-cell expression dataset [9]. Consequently, all cell types which are contained in the bulk mixtures should ideally also be included in the reference matrix [13]. This, however, represents a major issue in practical applications. Bulk profiles likely contain signals of cell types not originally considered for model development; for instance, hidden contributions from cancer cells likely do not share any similarity with the considered reference profiles.

Another, related issue is the origin of reference profiles. It is well known that the cell's molecular phenotype depends on its environment. This also affects the results of digital tissue dissection, as shown by [14, 15], where it was pointed out that the references should be selected from niches similar to the investigated bulks.

Although many different approaches were meanwhile suggested for cell-type deconvolution, they rarely account for hidden cell type contributions and cellular environmental effects on the reference profiles. Notably, in seminal work by [15], the tumor background was estimated together with the cellular proportions of a set of pre-specified cell types, and the different single-cell datasets were considered simultaneously to account for diversity between datasets [12]. Here, we systematically address both issues within a single approach named Adaptive Digital Tissue Deconvolution (ADTD). First, ADTD builds on our previous work on Digital Tissue Deconvolution (DTD) and uses gene weights for optimized deconvolution. Second, it augments the deconvolution by estimates of a background profile together with corresponding cellular proportions across a set of investigated bulks. Third, ADTD adapts its reference profiles to the investigated bulks. Intriguingly, the latter also resolves cell-type specific regulation from bulk transcriptomics data. ADTD was verified in

simulations based on single-cell RNA sequencing data of healthy and breast cancer tissue, and further, we demonstrate its application to resolve cell-type specific molecular differences between breast cancer subtypes.

2 Methods

Adaptive digital tissue deconvolution (ADTD) builds on loss-function learning for Digital Tissue Deconvolution (DTD) introduced previously by 10 and extends it by both accounting for unknown background contributions and adaptive reference profiles. We will first recapitulate the concept of loss-function learning and will then introduce ADTD.

2.1 Loss-function learning for digital tissue deconvolution

Bulk gene expression profiles can be modeled as a linear combination of cell-type-specific reference profiles,

$$Y = XC + \epsilon\,, \tag{1}$$

where $Y \in \mathbb{R}_+^{p\times n}$ is a matrix with n bulk gene expression profiles and $X \in \mathbb{R}_+^{p\times q}$ a matrix with q cell-type specific reference profiles in their columns, both containing expression values of p genes in their rows. The columns of $C \in \mathbb{R}_+^{q\times n}$ contain cellular weights corresponding to the columns (bulks) of Y. The values C_{ki} can be interpreted as semi-quantitative measurements of the number of cells of type k in bulk profile i. The residuals between the observed bulks Y and the fitted bulks $\hat{Y} = X\hat{C}$ are given by $\epsilon \in \mathbb{R}^{p\times n}$.

One possible way to estimate C is by

$$\hat{C} = \arg\min_C ||Y - XC||_F^2\,, \tag{2}$$

where $||.||_F$ is the Frobenius norm and $\hat{C}$ the estimate of C for given X and Y. Eq. (2) can be solved for each column (bulk) in Y, individually. In [10, 16], Eq. (2) was augmented by gene weights $g = (g_1, g_2, \ldots, g_p)^T$, yielding

$$\hat{C} = \arg\min_C ||\operatorname{diag}(g_1, g_2, \ldots, g_p)(Y - XC)||_F^2\,, \tag{3}$$

where the weights g_i can be chosen to improve estimates of C. They can be either specified by prior knowledge or can be learned from cellular mixtures with known cellular contributions, such as annotated bulk profiles or artificial mixtures generated from single-cell RNA sequencing data [10, 16]. To optimize the gene weights, we defined the outer loss function

$$L(g) = -\sum_{j=1}^{q} \operatorname{cor}(C_{j\cdot}, \hat{C}_{j\cdot}(g)) \quad \text{subject to } ||g||_2 = 1\,, \tag{4}$$

where $C_{j\cdot}$ is the jth row (cell type) of C and $\hat{C}_{j\cdot}$ its corresponding estimate. Thus, minimizing loss function Eq. (4) with respect to the g_is, maximizes the Pearson's correlation between the ground truths $C_{j\cdot}$ and its estimates $\hat{C}_{j\cdot}$ for all cell types of interest j, simultaneously. In contrast to the original implementation of loss-function learning [10], we additionally restricted the optimization problem to non-negative (physical) solutions $\hat{C} \succeq 0_{q,n}$. The corresponding algorithm uses the PyTorch machine learning library [17] with stochastic gradient descent for optimization.

2.2 Estimates of background contributions

One major drawback of reference-based cell-type deconvolution methods is that they do not necessarily capture all cellular contributions present in complex biospecimens. For instance, let X consist of a set of immune cell reference profiles and let Y be the bulk profile of a tumor specimen. Then, although immune cell contributions can be estimated, their values might be biased or blurred by the contributions from cancer cells to Y. In seminal work by 15, this issue was addressed by a method called EPIC, as outlined in the following. Let $Y_{\cdot i}$ be an individual bulk profile corresponding to the ith column of Y. Then we have

$$Y_{\cdot i} \quad = \quad XC_{\cdot i} + xc_i + \epsilon_{\cdot i}\,, \tag{5}$$

where we made the contribution from a potential background explicit by adding the term xc_i. Here, $x \in \mathbb{R}^{p\times 1}$ represents the "hidden" background profile and c_i its corresponding scalar weight. We further assume that the bulk profiles, the reference profiles, and the background profile are normalized by

$$\sum_{j=1}^{p} Y_{ji} = 1\,,\ \text{for } i = 1,\ldots,n\,; \quad \sum_{j=1}^{p} x_{j1} = 1\,; \tag{6}$$

$$\sum_{j=1}^{p} X_{jk} = 1\,,\ \text{for } k = 1,\ldots,q\,,$$

which gives $1 = \sum_{j=1}^{p} Y_{ji} = \sum_{j=1}^{p}\sum_{k=1}^{q} X_{jk}C_{ki} + \sum_{j=1}^{p} x_{j1}c_i + \sum_{j=1}^{p}\epsilon_{ji} = \sum_{k=1}^{q} C_{ki} + c_i + \epsilon_i$ with $\epsilon_i = \sum_{j=1}^{p}\epsilon_{ji}$. This implies $c_i = 1 - \sum_{k=1}^{q} C_{ki} - \epsilon_i$ and motivates that c_i can be directly estimated as

$$\hat{c}_i = \max\left\{0, 1 - \sum_{k=1}^{q}\hat{C}_{ki}\right\}\,. \tag{7}$$

Thus, EPIC intrinsically assumes that C is independent of the background profile x and that all ϵ_i are small.

2.3 Adaptive Digital Tissue Deconvolution

2.3.1 Background estimation

ADTD relaxes EPIC's assumptions and directly takes into account that the inference of C also depends on the background profile x. To achieve this, we replace the c_is in the residuals $\epsilon = Y - XC - xc$ of Eq. (5) by $c_i = 1 - \sum_{k=1}^{q} C_{ki}$ (assuming that the ϵ_is are small), and minimize their squared sum with respect to C and x. The corresponding optimization problem taking into account the gene weights determined via loss-function learning and non-negativity constraints becomes

$$\underset{C\succeq 0_{q,n},\, x\succeq 0_{p,1}}{\arg\min} \quad ||G(Y - XC - x(J_{1,n} - J_{1,q}C))||_F^2 + \lambda_1||C - C_0||_F^2\,, \tag{8}$$

subject to $J_{1,q}C \preceq J_{1,n}$ and $\sum_{j=1}^{p} x_{j1} = 1$, where $G = \mathrm{diag}(g_1, g_2, \ldots, g_p)$, C_0 is the estimate of C using an ordinary NNLS estimate with gene weights $\gamma_j = g_j^2$, and $J_{a,b}$ denotes a matrix of ones of dimension $a \times b$. The regularization parameter λ_1 calibrates between two solutions. For $\lambda_1 \to \infty$, the naive solution is retrieved, where the hidden cell proportions are given by Eq. (7). For $\lambda_1 \to 0$, x and C are estimated simultaneously. With growing λ_1 values, the solutions are stronger biased toward the naive solution $\hat{C} = C_0$. Note that the naive solution can be determined for each cellular

mixture individually, while the latter solution requires a sufficient amount of bulk samples, however, providing the advantage that a consensus background is learned across a set of bulk specimens. We solved the optimization problem Eq. (8) by iteratively optimizing with respect to C and x using quadratic programming, until convergence is reached.

2.3.2 Adaptive reference profiles

Reference profiles can be generated using external data from purified cellular mixtures or from single-cell data. However, cells adapt their molecular phenotype to their environment and as a consequence, global and static reference profiles might be inappropriate for most applications. ADTD dynamically adapts reference profiles to the specific application in order to improve inferred cellular contributions and to reveal potential regulatory effects on cell populations. This is achieved by replacing the reference matrix X by $\Delta \circ X$, where Δ is a $p \times q$ matrix of rescaling factors which are applied component-wise to X ($\circ$ is the Hadamard product). The full ADTD loss function is defined as

$$\begin{aligned} & L_{\mathrm{ADTD}}(C, x, \Delta) \\ = \; & ||G\left(Y - (\Delta \circ X)C - x(J_{1,n} - J_{1,p}(\Delta \circ X)C)\right)||_F^2 \\ & + \lambda_1 ||C - C_0||_F^2 + \lambda_2 ||J_{p,q} - \Delta||_F^2 \,, \end{aligned} \tag{9}$$

which is optimized with respect to C, x, and Δ, subject to the constraints

$$\begin{gathered} C \succeq 0_{q,n}; \quad J_{1,q}(\Delta \circ X)C \preceq J_{1,n}; \quad x \succeq 0_{p,1}; \\ \Delta \succeq 0_{p,q}; \quad \sum_{j=1}^{p} x_{j1} = 1 \,. \end{gathered} \tag{10}$$

In Eq. (9), we replaced $c = (c_1, c_2, \ldots, c_n) \in \mathbb{R}_+^{1\times n}$ by $J_{1,n} - J_{1,p}(\Delta \circ X)C$ as derived in Supplementary Section 1.1. The regularization term $\lambda_2 ||J_{p,q} - \Delta||_F^2$ controls overfitting with respect to the parameters Δ_{jk}. For $\lambda_2 \to \infty$, we obtain $\Delta = J_{p,q}$.

2.3.3 Implementation

The ADTD optimization problem was solved by iteratively optimizing the loss $L_{\mathrm{ADTD}}(C, x, \Delta)$ with respect to C, x, and Δ until convergence is reached. The minimization with respect to C and x was addressed by quadratic programming as derived in Supplementary Materials Sections 1.2 and 1.3. The optimization with respect to Δ was performed iteratively for the rows $\Delta_{j\cdot}$. The latter solution was determined as outlined in Supplementary Section 1.4. The full algorithm is summarized below (Algorithm 1).

2.3.4 Hidden-cell simulation study

Training data We generated artificial mixtures of known cellular compositions for model training using single-cell data from healthy breast tissue (GSE164898) [18], representing women of different race, age, parity, menstrual phase, Tyrer-Cuzick score, and Body Mass Index (BMI). We denote these data as "training data" throughout the article. Cell types were labeled using Seurat [**?**]. The reference matrix X with prototypic reference profiles of B-cells, endothelial cells, myeloid cells, epithelial cells, perivascular-like cells (PVL), and T-cells were generated by averaging across all respective single-cell profiles of individual patients. This yielded a reference matrix for each patient. The matrices were averaged to yield a global reference matrix. Cells which were not labeled as one

Algorithm 1 Adaptive Digital Tissue Deconvolution

procedure ADTD$(X, Y, \gamma, \lambda_1, \lambda_2)$

  Normalize X and Y such that $\sum_j X_{jk} = 1$ and $\sum_j Y_{ji} = 1$

  Initialize $C = C_0$, $\Delta = J_{p,q}$, and $x = x_0$, and

  for $iteration = 1, 2, \ldots, i_{max}$ **do**
    update C (see Suppl. Section 1.2)
    update x (see Suppl. Section 1.3)
    for $k = 1, 2, \ldots, p$ **do**
      update $\Delta_{k\cdot}$ (see Suppl. Section 1.4)
    Terminate when desired precision is achieved

of these six cell types were removed from the training set. We further restricted the gene space to the 1000 most variable genes with respect to X. Individual mixtures ($n = 1000$) were generated by randomly drawing 100 single-cell profiles patient-wise and by averaging their expression.

Validation data Artificial validation mixtures of known cellular compositions were generated analogously to the training data, but now using single-cell data of breast cancer specimens [19]. In total, single-cell data of 26 different tumors were available with corresponding cell-type labels. We extracted B-cells, endothelial cells, myeloid cells, epithelial cells, PVL cells, and T-cells, as described above, and additionally cancer epithelial cells. These profiles were used to generate artificial bulk mixtures by drawing 400 single-cell profiles corresponding to individual samples, as previously for the training data.

2.3.5 Cellular regulation simulation study

To study ADTD's capability to reconstruct cell-type specific regulation from bulk transcriptomics data, we re-generated the validation data as above, but now artificially modified gene-expression levels of T-cells (corresponding to $\sim 39.1\%$ of the single-cell profiles). For this, we modified the expression of gene j in mixture i by (1) drawing a modification factor a_i from the set $\{-1, -0.5, 1, 2\}$, and (2) by adding $a_i X_{kj} C_{ki}$ to the original expression Y_{ij} (values $Y_{ij} < 0$ were set to zero). In each simulation run, we modified 20 genes; 10 T-cell unspecific genes were selected according to minimal values of $\mathrm{var}(X_{j\cdot})/\,\mathrm{mean}(X_{j\cdot})$ and 10 T-cell specific genes selected via maximal $X_{jk}/\,\mathrm{mean}(X_{j\cdot})$. For an alternative analysis, we artificially modified the gene expression in PVL cells ($\sim 6.0\%$ of single-cell profiles) following the same strategy as outlined for T cells.

3 Results

3.1 ADTD recovers hidden cellular contributions

In this section, we show that ADTD facilitates the reconstruction of hidden cellular contributions in simulation studies. We point out the critical role of prior gene selection and how it affects ADTD's and competing methods' performance. We further show that ADTD's estimates are more accurate both in a scenario with well-selected marker genes and in a scenario where the marker genes insufficiently represent the hidden contributions.

3.1.1 Marker-gene selection and weighting

It was repeatedly shown that gene selection strongly affects the performance of bulk deconvolution methods [15, 10], and the importance of a stringent marker gene selection was thoroughly demonstrated [13]. To identify and weigh genes according to their importance for tissue dissection, we used artificial mixtures of single-cell RNA sequencing profiles from healthy breast tissue (see Methods). These training data were used for loss-function learning to determine gene weights g_j and to generate a gene importance ranking (ordered by $g_j^2/\operatorname{var}(X_{j\cdot})$) to disentangle B-cells, endothelial cells, myeloid cells, normal epithelial cells, PVLs, and T-cells. The weights were used for ADTD and the gene ranking was used to build an appropriate reference matrix for EPIC. We evaluated two versions of ADTD, one strongly regularized version ($\lambda_1 \to \infty$), which can also be applied to individual samples, and a version where a consensus background is estimated together with all cellular proportions ($\lambda_1 = 10^{-5}$, see also Suppl. Methods Section 2 for details about hyper-parameter selection). The former will be denoted as ADTD_1 and the latter as ADTD_2 in the following. For pedagogical reasons, we first set the second regularization parameter λ_2 to infinity, which means that no adaptation of reference profiles is inferred. We compared $\text{ADTD}_{1,2}$ to two different implementations of EPIC. For EPIC_1, we evaluated the training performance using the top $p = 100, 200, \ldots, 1000$ genes as provided by loss-function learning and selected the best-performing one. For EPIC_2, we followed the same strategy but now used the reference matrix as provided in the EPIC *R* package 20, which, however, returns only estimates for B-cells, endothelial cells, macrophages, PVLs, and T-cells. Macrophages estimated by EPIC_2 were compared to the ground truth of myeloid lineage cells. Note that many new approaches for cell-type deconvolution were suggested recently [8, 9, 11, 12, 4, 5], which, however, are not able to estimate background contributions. Therefore, we restricted the following comparisons to EPIC, since it provides estimates of background contributions and since it was shown to perform well in comparison to other available methods with respect to the correlation between ground truth proportions and their estimates [13]. Note also that it becomes increasingly clear that the employed deconvolution method is only one of many factors which determine model performance. As thoroughly investigated in [13], data normalization and particularly gene selection turned out to be crucial. The training performance of ADTD_1, ADTD_2, EPIC_1, and EPIC_2 for each of the investigated cell types is shown in Table 1. We observed that ADTD_1 performs best with an average Spearman correlation between estimated cellular proportions and the corresponding ground truth of 0.915, with cell-type specific performance ranging from 0.63 (B-cells) to 0.99 (T-cells and epithelial cells). These results are followed by ADTD_2 with an average performance of 0.807, EPIC_1 with 0.783, and EPIC_2 with 0.492. The ADTD training performance for different choices of regularization parameters λ_1 is summarized in Table S1, and for different sample sizes n in Figure S2.

3.1.2 Reconstruction of hidden cellular contributions using well selected marker genes

Next, we validated the performance on the external validation data, where we now also considered the influence of background contributions from both normal and cancer epithelial cells, contributing, on average, $\sim 31\%$ to the artificial mixtures ($\sim 4.7\%$ epithelial and $\sim 26\%$ cancer epithelial cells, respectively). First, we applied $\text{ADTD}_{1,2}$ and $\text{EPIC}_{1,2}$ to $n = 1000$ validation mixtures, where we removed the reference profile for normal epithelial cells; there is now neither a reference for normal nor cancer epithelial cells present. We refer to this study as validation scenario A throughout this article. Corresponding results are shown in Table 2. We observed that both ADTD approaches perform substantially better than EPIC_1, which is particularly true for the hidden contribution consisting of epithelial and cancer epithelial cells. EPIC_2 showed a performance comparable to $\text{ADTD}_{1,2}$ in estimating B-cells and myeloid cells, while T-cells and endothelial cells could no longer be

Table 1: **ADTD and EPIC training performance.** Observed Spearman correlations obtained by comparing the estimated cellular proportions with the ground truth for artificial cellular training mixtures generated from single-cell data of healthy breast tissue specimens (see Methods). The errors correspond to ±1 standard deviation obtained over 10 simulation runs. Abbreviations: Endo. = endothelial cells; Myel. = myeloid cells; Epith. = epithelial cells; PVL = perivascular-like cells

	$ADTD_1$	$ADTD_2$	$EPIC_1$	$EPIC_2$
B-cells	0.629 ± 0.023	0.39 ± 0.085	0.303 ± 0.018	0.215 ± 0.016
Endo.	0.984 ± 0.002	0.858 ± 0.017	0.854 ± 0.011	0.756 ± 0.01
Myel.	0.982 ± 0.001	0.809 ± 0.029	0.898 ± 0.007	0.76 ± 0.02
Epith.	0.986 ± 0.001	0.974 ± 0.04	0.949 ± 0.005	-
PVL	0.923 ± 0.004	0.857 ± 0.024	0.769 ± 0.008	-
T-cells	0.986 ± 0.001	0.955 ± 0.013	0.927 ± 0.006	0.236 ± 0.027
mean	0.915 ± 0.004	0.807 ± 0.02	0.783 ± 0.005	0.492 ± 0.007

Table 2: **ADTD and EPIC performance in validation study *A*.** Observed Spearman correlations obtained by comparing the estimated cellular proportions with the ground truth for artificial cellular validation mixtures, where both epithelial and cancer epithelial cells were hidden in the mixtures and not represented as reference profile. The epithelial background contributions (epithelial + cancer epithelial) is highlighted in boldface. The errors correspond to ±1 standard deviation obtained over 10 simulation runs. Abbreviations: Myel. = myeloid cells; Epith. = epithelial cells; PVL = perivascular-like cells

	$ADTD_1$	$ADTD_2$	$EPIC_1$	$EPIC_2$
B-cells	0.727 ± 0.011	0.747 ± 0.015	0.753 ± 0.017	0.827 ± 0.011
Endo.	0.854 ± 0.01	0.834 ± 0.014	0.556 ± 0.022	0.133 ± 0.028
Myel.	0.873 ± 0.009	0.847 ± 0.01	0.851 ± 0.005	0.773 ± 0.008
PVL	0.737 ± 0.01	0.862 ± 0.006	0.86 ± 0.011	-
T-cells	0.814 ± 0.006	0.875 ± 0.005	0.726 ± 0.016	-0.044 ± 0.017
hidden Epith.	**0.942 ± 0.005**	**0.931 ± 0.01**	**0.673 ± 0.015**	**-**
mean	0.824 ± 0.003	0.849 ± 0.003	0.736 ± 0.005	0.422 ± 0.008

reliably estimated. $EPIC_2$ background contributions were not assessed since they are a mixture of all remaining cells, hindering the performance comparison. Moreover, we observed better performance of $ADTD_2$ then for $ADTD_1$ in contrast to the training performance. Note that $ADTD_2$ is constructed such that it can better deal with background contributions than $ADTD_1$, which naturally explains the improved performance if such a background is present. This also becomes apparent when the estimates of the "hidden" reference profile x are compared to its ground truth (Spearman $\rho \sim 0.581 \pm 0.006$ for $ADTD_1$ and $\rho \sim 0.93 \pm 0.002$ for $ADTD_2$, respectively). Note that the ground truth reference profile is now a combination of both normal and cancer epithelial cells. Importantly, the gene selection/weighting performed previously identified genes to disentangle epithelial from other cells, and as such contains information to reconstruct the hidden background (which is a mixture of healthy and cancerous epithelial cells).

Sample-size dependency $ADTD_2$ solves an optimization problem which simultaneously estimates the hidden reference profile x and its cellular proportion. Thus, an accurate estimate of x requires accurate estimates of the known contributions C and vice versa. The hidden background

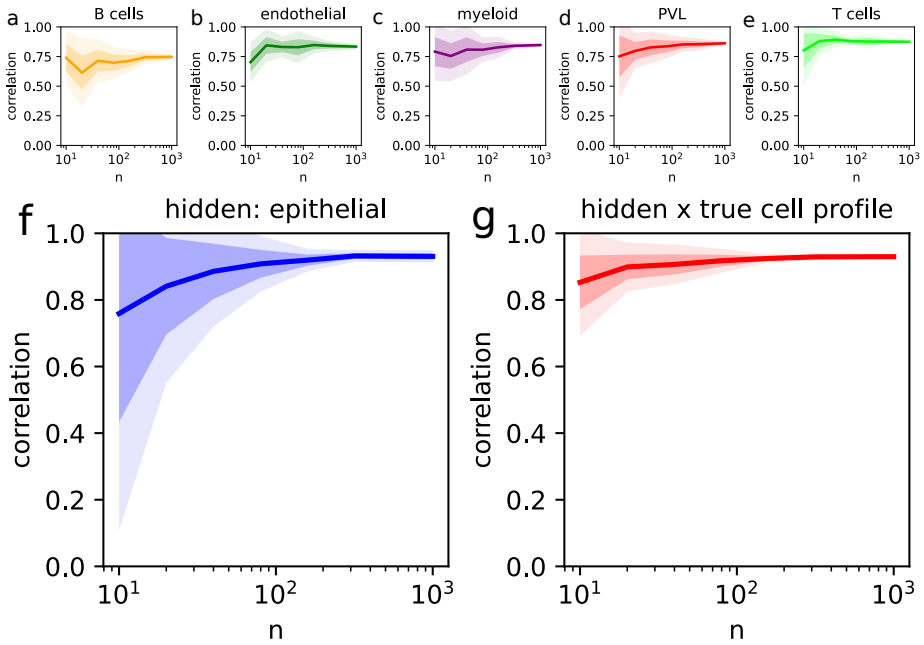

Figure 1: **Simulation illustrating sample-size stability of ADTD$_2$ in validation scenario *A*.** Performance versus sample size n for ADTD$_2$ in terms of Spearman's rank correlation for comparing estimates with their ground truth. Figures a to e show the performance for cell types captured in the reference matrix X, corresponding to B-cells, endothelial cells, myeloid cells, PVLs, and T-cells, respectively. Figure f shows the corresponding n dependency for the hidden proportion, which is here a combination of epithelial and cancer epithelial cells. Figure g shows the Spearman's correlation for comparing the estimated background profile x with the ground truth. The dark shaded bands correspond to ± 1 standard deviation and the light shaded bands to ± 2 standard deviations calculated across 10 simulation runs.

profile x can be considered as an average background which is observed among different mixtures. Thus, to obtain reliable estimates for both x and c, one might need to consider multiple mixtures. To study the impact of sample size n on ADTD$_2$ performance, we repeated the previous validation study for different numbers of samples with results shown in Figure 1, where the observed Spearman's rank correlation is plotted versus sample size n for the cell types contained in the reference matrix (Figs. 1a to e), and for the estimated hidden proportions (Fig. 1f). Additionally, Fig. 1g shows the Spearman's correlation comparing the estimated background profile x with the ground truth. Strikingly, ADTD$_2$ achieves already remarkable performance for sample sizes as low as $n = 10$. A corresponding study evaluating the training data for ADTD$_{1,2}$, yielded the results shown in Supplementary Figure S2, where, however, the artificial mixtures did not contain background contributions and thus were not evaluated.

Table 3: **ADTD and EPIC performance in validation study B.** Observed Spearman correlations obtained by comparing the estimated cellular proportions with the ground truth for artificial cellular validation mixtures, where solely the cancer epithelial cells were hidden in the mixtures and not represented as reference profile (highlighted in boldface). The errors correspond to ±1 standard deviation obtained over 10 simulation runs. Abbreviations: Myel. = myeloid cells; Epith. = epithelial cells; PVL = perivascular-like cells

	$ADTD_1$	$ADTD_2$	$EPIC_1$	$EPIC_2$
B-cells	0.661 ± 0.018	0.732 ± 0.023	0.7 ± 0.015	0.827 ± 0.011
Endo.	0.803 ± 0.016	0.735 ± 0.019	0.648 ± 0.014	0.133 ± 0.028
Myel.	0.841 ± 0.012	0.787 ± 0.015	0.864 ± 0.006	0.773 ± 0.008
Epith.	0.1 ± 0.017	0.139 ± 0.014	0.123 ± 0.013	-
PVL	0.875 ± 0.01	0.811 ± 0.014	0.837 ± 0.01	-
T-cells	0.888 ± 0.005	0.908 ± 0.007	0.938 ± 0.003	-0.044 ± 0.017
hidden cancer Epith.	**0.208 ± 0.013**	**0.648 ± 0.017**	**-0.13 ± 0.016**	**-**
mean	0.625 ± 0.004	0.68 ± 0.006	0.569 ± 0.005	0.422 ± 0.008

3.1.3 Reconstruction of hidden cellular contributions with non-optimal marker gene selection

In practical applications, it cannot be guaranteed that all cellular contributions including a possibly hidden background are sufficiently represented by the marker gene selection or weighting. Therefore, we studied non-optimal marker gene selection in the validation data as follows and refer to this study as validation scenario B. We retained the full reference matrix X as applied for the training procedure and generated artificial mixtures analogously to the previous validation study including cancer epithelial contributions. Then, in contrast, we evaluated the background reconstruction of solely the hidden cancer epithelial cells and not the mixture of both epithelial and cancer epithelial cells. Therefore, a background is reconstructed for which the marker gene selection was not optimized. We obtained the results shown in Table 3, where $ADTD_2$ now substantially outperforms $ADTD_1$ with respect to the reconstruction of the hidden background profile x (Spearman $\rho = 0.28 \pm 0.008$ vs. $\rho = 0.62 \pm 0.08$ for $ADTD_1$ vs. $ADTD_2$) and its corresponding proportions (Spearman $\rho = 0.208 \pm 0.013$ vs. $\rho = 0.648 \pm 0.017$ for $ADTD_1$ vs. $ADTD_2$). Further, we evaluated $EPIC_1$, showing clearly inferior performance compared to $ADTD_2$, and performance comparable to $ADTD_1$. $EPIC_2$ background contributions were not evaluated for the reasons outlined above. All methods were not able to correctly estimate healthy epithelial cells with the best performance observed for $ADTD_2$ with $\rho = 0.139 \pm 0.014$. One should be aware that the definition of healthy epithelial cells in the training data might substantially deviate from those in the validation data due to different strategies for cell labeling (Seurat vs. genomic instability score).

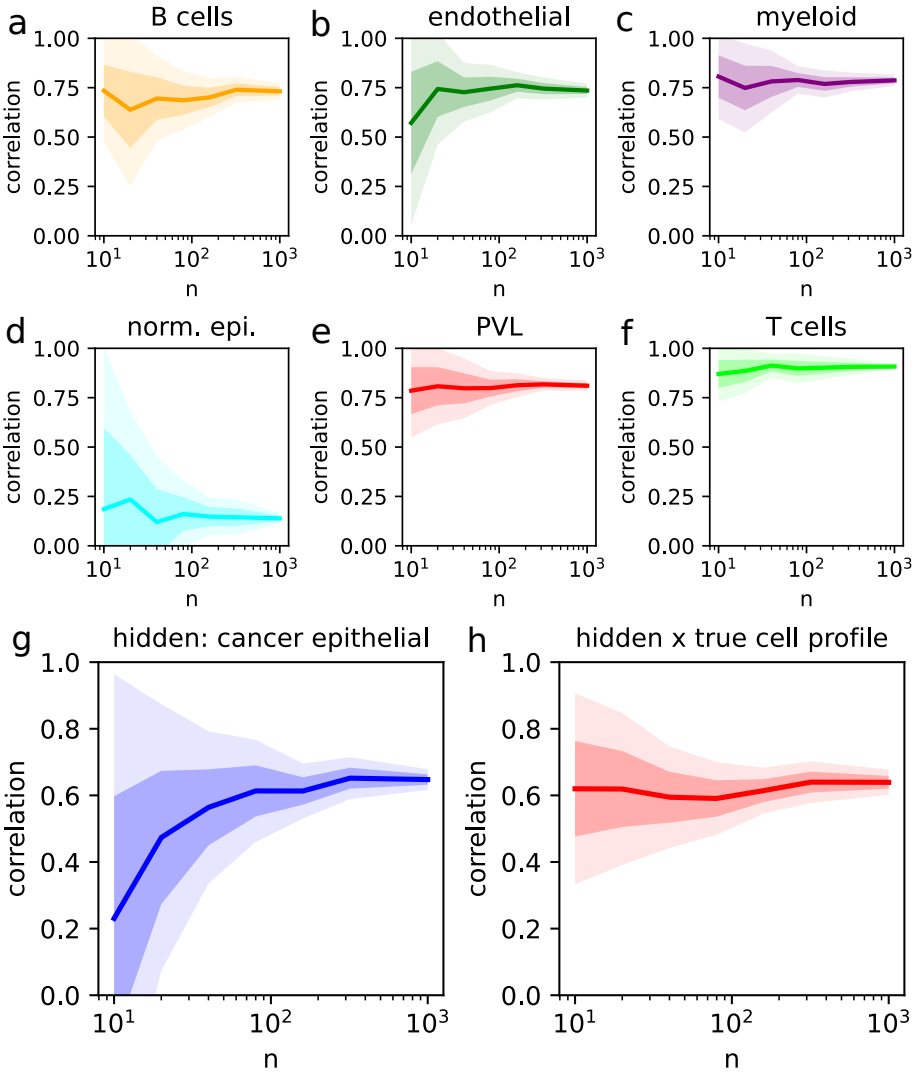

Figure 2: **Simulation illustrating sample-size stability of ADTD$_2$ in validation scenario B.** Performance versus sample size n for ADTD$_2$ in terms of Spearman's rank correlation for comparing estimates with their ground truth. Figures a to e show the performance for cell types captured in the reference matrix X, corresponding to B-cells, endothelial cells, myeloid cells, normal epithelial, PVLs, and T-cells, respectively. Figure f shows the corresponding n dependency for the hidden proportion, which is here given by only cancer epithelial cells. Figure g shows the Spearman's correlation for comparing the estimated background profile x of cancer epithelial cells with the ground truth. The dark shaded bands correspond to ± 1 standard deviation and the light shaded bands to ± 2 standard deviations calculated across 10 simulation runs.

Sample-size dependency We evaluated the sample-size dependency of ADTD_2 as in the previous simulation study with results shown in Figure 2. Also with non-optimal marker gene selection, we observed qualitatively the same results; the reconstruction of both the known and unknown, hidden background contributions is possible even for low sample sizes. Here, however, we observed that very low sample sizes $n \sim 10 - 50$ yielded less reliable results for the estimated background proportions compared to $n \geq 100$.

3.2 Reverse engineering of cellular regulation in breast cancer

ADTD allows estimating both cellular background and cellular regulation. In previous simulations, we only estimated the former and ignored the latter which was achieved by $\lambda_2 \to \infty$. Here, we illustrate that ADTD can resolve both simultaneously. For this, we repeated the two simulation studies outlined above, but now modified selected gene-expression levels in T-cells (see Methods) and applied $\text{ADTD}_{1,2}$ for different regularization parameters $\lambda_2 = 10^{-5}, 10^{-4}, \ldots, 0.1, 1$. Considering the more benign scenario where genes were selected to also capture background contributions (validation scenario A), we evaluated the Receiver Operating Characteristic (ROC) curve for detecting modified genes in T-cells with corresponding areas under the curve (AUCs) shown in Figure 3a and b for different regularization parameters, where Fig. 3a corresponds to ADTD_1 ($\lambda_1 \to \infty$) and Fig. 3b to ADTD_2 ($\lambda_1 = 10^{-5}$). We observed that ADTD_1 yields stable AUCs of ~ 0.7 except for the very small regularization of $\lambda_2 = 10^{-5}$, where the performance was strongly compromised. The more flexible implementation of ADTD_2 shows a completely different behavior. While modified genes are detected with similar AUCs for large regularization parameters $\lambda_2 = 10^{-2}, 0.1, 1$, the performance improves for the low values $\lambda_2 = 10^{-5}, 10^{-4}, 10^{-3}$ with best performance observed for $\lambda_2 = 10^{-5}$. Notably, since ADTD estimates cellular proportions, background contributions, and cellular regulation jointly, we also investigated how the choice of λ_2 affects the performance in reconstructing the former two with results shown in Fig. 3c and d for ADTD_1 and ADTD_2, respectively. ADTD_1 shows a perfectly stable performance with respect to Spearman's correlation between estimates and ground truth for both the covered and hidden contributions, while the performance of ADTD_2 depends on λ_2 with improved performance for hidden epithelial and cancer cells and decreased performance for B-cells for lower λ_2 values. The compromised performance of B-cells might be explained by their small average abundance of $\sim 0.4\%$ in the training data ($\sim 3.6\%$ in the validation data).

A corresponding simulation study considering validation scenario B (non-optimal gene selection) is shown in Figure 4a and b, where AUCs are contrasted with λ_2, and in Figure 4c and d, where the corresponding performance in terms of Spearman's correlations is provided. The results qualitatively agree with those obtained in validation scenario A.

To further substantiate our findings, we manipulated gene expression levels in PVLs, which only contributed a minor proportion to the cellular mixtures of $\sim 6.9\%$. Remarkably, the performance was only little affected and both cellular regulation and cellular proportions of both known and hidden contributions could be estimated with confidence (see Supplementary Figure S3 and S4 for validation study A and B, respectively).

Finally, we studied the effect of sample size. For this purpose, we re-evaluated the models ADTD_1 ($\lambda_1 \to \infty$, $\lambda_2 = 10^{-4}$) and ADTD_2 ($\lambda_1 = 10^{-5}$, $\lambda_2 = 10^{-4}$) in validation scenario A and B for different sample sizes $n = 125, 250, 500, 1000$ in Supplementary Figure S5 and S6 (manipulated T-cells) and S7 and S8 (manipulated PVLs). We obtained slightly compromised performance for the lower sample sizes (see Figs. S6 and S8, validation scenario B) or approximately equal performance (see Figs. S5 and S7, validation scenario A).

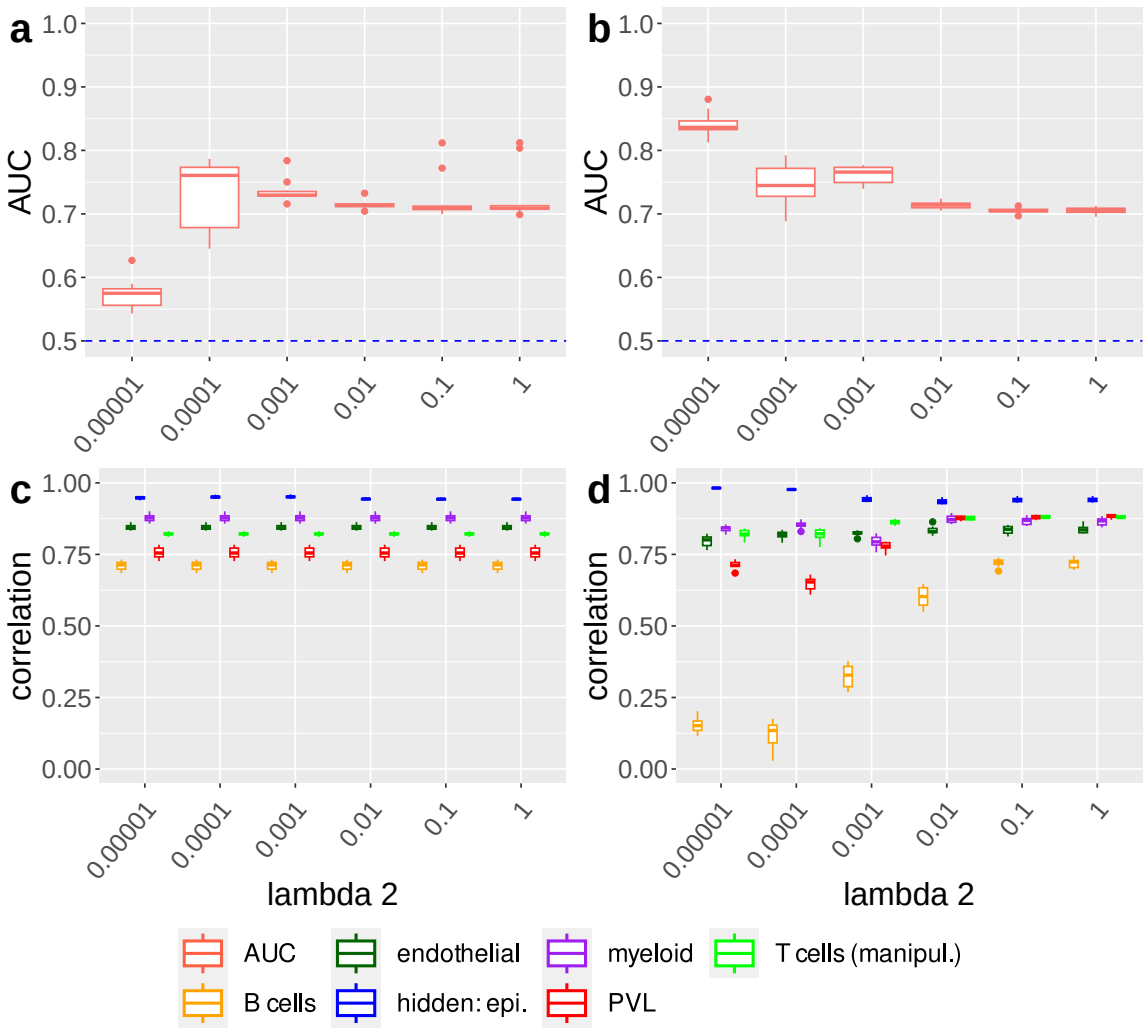

Figure 3: **ADTD recovers cellular regulation and retains prediction performance in validation scenario** A**.**The upper row shows ROC AUCs for recovering cellular regulation simulated in T-cells for different regularization parameters λ_2 for ADTD_1 (left) and ADTD_2 (right) in validation study A. The lower row shows the corresponding performance in terms of Spearman's correlation for ADTD_1 (left) and ADTD_2 (right) for estimating the known and hidden cellular contributions. Abbreviation: "hidden epi." = hidden epithelial cells (healthy + cancer epithelial)

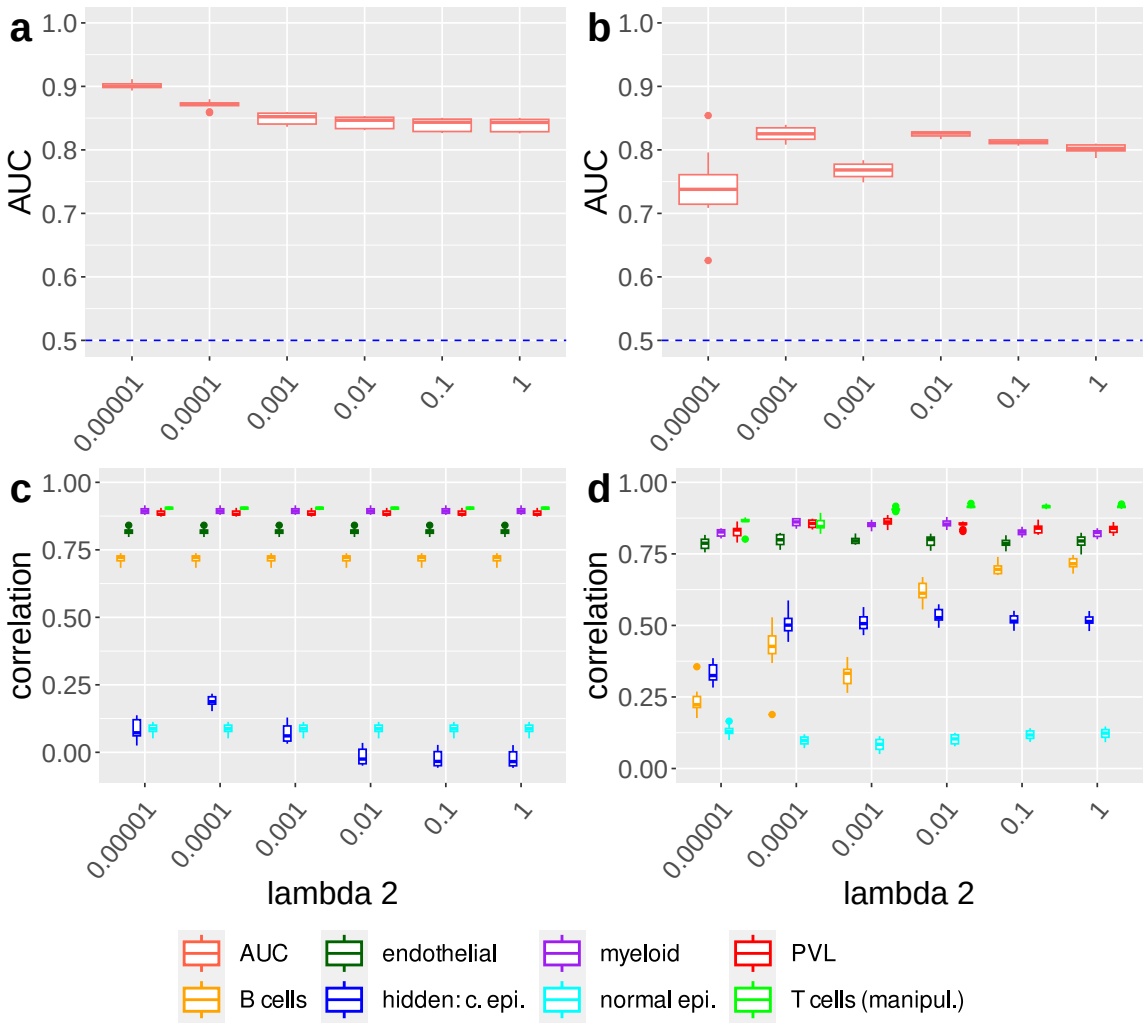

Figure 4: **ADTD recovers cellular regulation and retains prediction performance in validation scenario** A**.**The upper row shows ROC AUCs for recovering cellular regulation simulated in T-cells for different regularization parameters λ_2 for ADTD_1 (left) and ADTD_2 (right) in validation study B. The lower row shows the corresponding performance in terms of Spearman's correlation for ADTD_1 (left) and ADTD_2 (right) for estimating the known and hidden cellular contributions. Abbreviation: "hidden: c. epi." = hidden cancer epithelial cells

3.3 ADTD breast cancer sub-type analysis

To illustrate how ADTD can complement routine data analysis, we performed the following experiment. We generated data as previously in validation scenario B for in total $n = 1000$ artificial bulk samples ($p = 1000$) for each of the three breast cancer subtypes, estrogen receptor-positive breast cancer (ER+), human epidermal growth factor receptor 2-positive breast cancer (HER2+), and triple-negative breast cancers (TNBC). The following analysis requires only a labelling of the cancer subtype and information from the training procedure; a reference matrix (containing B-cells, endothelial cells, myeloid cells, epithelial cells, PVL, and T-cells), gene weights and a reasonable set of hyper-parameters. Importantly, this analysis does not require any cell-type labelling of the single-cell profiles used to generate the artificial bulk samples, meaning that ADTD evaluates groups of samples relative to the healthy breast specimens used for model development. As such, this example analysis can be considered as complementary to a standard single-cell analysis. The hyper-parameters assessment in Supplementary Section 2 suggest a wide parameter range of λ_1 and λ_2 values with good performance on the training data, where we finally selected $\lambda_1 = 10^{-5}$ and $\lambda_1 = 10^{-4}$. Although this specific parameter set did not perform particularly well compared to the others, it allows to adapt the model rather strongly to the data due to the parameter's relatively small size. ADTD then yielded a parameter matrix Δ for each of the subtypes. First, we ranked the rows (genes) of Δ from highest to lowest deviations from one defined by $\max_k |\Delta_{jk} - 1|$. Thus, we identified the genes with strongest regulation according to ADTD for each of the three cancer subtypes across all cell types with included reference profile. In a first instance, we extracted the 20 top hits for each subtype and evaluated the overlap. This yielded four genes included in all three rankings (CITED4, CSN2, EMP2, MIF), eight genes in two of them, and 32 which just appeared in one of the three rankings. Strikingly, for each of the three, CITED4, CSN2, and EMP2, ADTD identified the same cell-type strongly regulated into the same direction across the three cancer subtypes. For instance, CITED4 was strongly down-regulated in normal epithelial cells consistently in all three subtypes ($\Delta_{\text{ER+,Epi.}} = 0.11$, $\Delta_{\text{HER+,Epi.}} = 0.20$, and $\Delta_{\text{TNBC,Epi.}} = 0.11$). The protein encoded by CITED4 is known to act as a transcriptional coactivator for TFAP2/AP-2 [21], and it was shown that breast cancer development is characterized by either nuclear loss or cytoplasmic translocation of CITED4 [22]. Similarly, casein Beta (CSN2) was down-regulated in normal epithelial cells consistently in all three subtypes ($\Delta_{\text{ER+,Epi.}} = 0.05$, $\Delta_{\text{HER+,Epi.}} = 0.22$, and $\Delta_{\text{TNBC,Epi.}} = 0.15$), but at slightly different rates. Casein proteins are secreted in human milk and human breast tumors lack beta-casein production [23]. In contrast, EMP2, which encodes the epithelial membrane Protein 2, was consistently up-regulated in all three subtypes again in healthy epithelial cells ($\Delta_{\text{ER+,Epi.}} = 2.04$, $\Delta_{\text{HER+,Epi.}} = 2.09$, and $\Delta_{\text{TNBC,Epi.}} = 2.06$) and at almost identical rates. EMP2 was recently identified as a novel regulator of stemness in breast cancer cells, with its upregulation promoting tumor growth in vivo [24]. In contrast to the former three, the Macrophage Migration Inhibitory Factor (MIF) did not show a consistent pattern across the three subtypes with the highest modification attributed to B-cells in ER+ ($\Delta_{\text{ER+,B-cells}} = 2.62$), to myeloid cells in HER2+ ($\Delta_{\text{HER2+,Myel.}} = 0.28$), and to T-cells in TNBC ($\Delta_{\text{TNBC,T-cells}} = 2.07$).

As a further consistency check, we ranked the Δs jointly, where we ranked from highest to lowest regulation across all subtypes ($\max_{k,\Delta} |\Delta_{jk} - 1|$). This yielded the overall strongest regulation in gene PYCARD; in both the ER+ and HER+ subtype, we observed a strong up-regulation in normal epithelial cells ($\Delta_{\text{ER+,Epi.}} = 2.78$, $\Delta_{\text{HER+,Epi.}} = 3.10$), while there was no substantial regulation in the remaining cell types. This effect was not observed for TNBCs, where the healthy epithelial cells were only negligibly corrected ($\Delta_{\text{TNBC,Epi.}} = 1.05$). PYCARD, also known as ASC (Apoptosis-associated speck-like protein containing a CARD) functions as a mediator of apoptosis and inflammation and is expressed in cells undergoing apoptosis. PYCARD was followed by MIF, which was already iden-

tified and discussed above, and by ACKR1 and IL7R. ACKR1 (Atypical Chemokine Receptor 1 (Duffy Blood Group)) was strongly regulated in endothelial cells but with very different strength in the three subtypes ($\Delta_{\text{ER+,Endo.}} = 2.59$, $\Delta_{\text{HER+,Endo.}} = 1.15$, and $\Delta_{\text{TNBC,Endo.}} = 1.52$). ACKR1 encodes a glycosylated membrane protein, which is a non-specific receptor for several chemokines. It was suggested that targeting ACKR1 on the tumor vasculature may prove efficacious in slowing or reversing tumor growth [25]. Interleukin receptor 7 (IL7R), in contrast, was strongly regulated in T-cells, but only in the ER+ subtype and slightly in the HER2+ subtype ($\Delta_{\text{ER+,T-cells}} = 2.42$, $\Delta_{\text{HER+,T-cells}} = 1.41$, and $\Delta_{\text{TNBC,T-cells}} = 1.04$). IL7R signalling is involved in multiple stages of lymphocyte development and maturation, while it is unnecessary in other stages [26]. Specifically, it is involved in V(D)J recombination of the T-cell receptor β gene [27]. The outlined findings indicate a phenotype in healthy epithelial cells which already resembles that of cancer epithelial cells, suggest that endothelial cells enable tumor angiogenesis cancer sub-type specifically, and provide hints towards an immune reaction in ER+ breast cancers.

4 Conclusion and Discussion

Multiple methods for digital tissue dissection were suggested in recent years, based on different losses such as squared residuals, l_1 or the support vector regression [2, 1]. Moreover, machine learning techniques such as deep learning were proposed as an alternative and promise improved performance, although strict Standard Operating Procedures (SOPs) might be necessary to achieve such a performance gain [5, 4]. Here, we approached computational tissue dissection from the perspective that strict SOPs are difficult (if not impossible) to guarantee and that methods should intrinsically account for between-dataset variability. As such, we proposed Adaptive Digital Tissue Deconvolution (ADTD) as an adaptive method for digital tissue dissection, which incorporates two sources of between-dataset variability. First, ADTD accounts for possibly hidden cellular background contributions, which are not represented by prototypical reference profiles. This allowed us to achieve improved out-of-sample performance as demonstrated by artificial validation mixtures generated from single-cell RNA profiles of breast cancer specimens. In fact, both the proportions of background contributions and their representation as a reference profile could be estimated with confidence even for low sample sizes of $n \sim 100$. Second, ADTD accounts for environmental effects; depending on the tissue context, the appearance of the prototypical reference profiles might be altered [14, 15] and therefore ADTD adapts its reference matrix to the specific application. As demonstrated in comprehensive simulation studies, this allows to resolve cell-type specific regulation from bulk transcriptomics data.

It is worth emphasizing that approaching the former two issues requires appropriate strategies to deal with overfitting. Here, a common technique from high-dimensional statistics was used, namely l_2 regularization [28]. This choice is not unique and was done for computational reasons since ADTD Eq. (9) can be minimized by iteratively solving sub-problems using quadratic programming. In principle, l_1 regularization or other sparseness-inducing regularization terms could be an interesting alternative to ADTD, which, however, requires appropriate implementation strategies. It is also important to point out that tuning the strength of the two l_2-regularization terms is achieved by adjusting two hyper-parameters λ_1 and λ_2, and selecting these parameters is highly non-trivial as outlined in the following. First, regularization parameters are usually selected such that the generalization error is as small as possible. This requires estimates of the latter. However, the formal objective of ADTD is to minimize the squared residuals between the observed bulk profiles Y and their reconstruction $\hat{Y}$. ADTD's practical objective is to estimate model parameters representing the cellular proportions (C), the background contributions (x and c), and the adaptation terms (Δ)

as reliable as possible. We can only assess the former generalization error on external data using techniques such as Cross Validation (CV) but not the latter, which we would be more interested in. Second, we can in principle estimate the generalization error of model parameters on the training data via CV using artificial cellular mixtures. However, cell-type deconvolution methods are usually applied to data which do not perfectly fit the original training data used for gene selection and weighting. As emphasized repeatedly, new data might contain additional hidden contributions or might differ due to environmental effects. Thus, even if we have an estimate of the generalization error, it might not reflect the true generalization error if applied to new data. Therefore, we pursued the following strategy. On the training data, we first evaluated ADTD for different choices of λ_1 with $\lambda_2 \to \infty$ fixed (regularizing the background contribution only) and selected the model with the smallest λ_1 value which still provided acceptable performance on the training data. Thus, we selected the model which still works on the training data (even though with compromised performance) but which accounts for as much between-dataset variability as possible. The same strategy was subsequently followed for determining λ_2. In fact, this yielded excellent performance in both validation studies A and B with respect to both estimates of background contributions and cell-type specific regulation. Anyway, more systematic approaches to determine appropriate values for λ_1 and λ_2 would reduce user-dependent choices. Possible directions for future research could be the adaptation of information criteria such as the Bayes' Information Criterion and extensions thereof [29]. This, however, would require sparseness constraints such as l_1 regularization, which are beyond the scope of this article. We would also like to point the reader towards recent research in machine learning where similarly to digital tissue dissection, the generalization error cannot be reliably estimated. This includes problems of unsupervised domain adaptation if data from a new domain is unlabeled [30], and counterfactual reasoning, where the counterfactual truth is generally not available [31], to name just two examples. As such, evaluation strategies are needed in general and not just in the context of digital tissue dissection.

Finally, ADTD represents a substantial step in the development of digital tissue dissection methods, as it is to the best of our knowledge the first method which takes into account two big intrinsic model uncertainties, namely hidden background contributions and cellular environmental factors.

5 Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [AL 2355/1-1 "Digital Tissue Deconvolution - Aus Einzelzelldaten lernen"]. SNG and FG acknowledge European Union's Horizon 2020 research and innovation programme (MESI-STRAT) [754688] and Trond Mohn Stiftelse [BFS2017TMT01].

Conflict of Interest: none declared.

References

[1] Binbin Chen, Michael S Khodadoust, Chih Long Liu, Aaron M Newman, and Ash A Alizadeh. Profiling tumor infiltrating immune cells with cibersort. In *Cancer systems biology*, pages 243–259. Springer, 2018.

[2] Zeev Altboum, Yael Steuerman, Eyal David, Zohar Barnett-Itzhaki, Liran Valadarsky, Hadas Keren-Shaul, Tal Meningher, Ella Mendelson, Michal Mandelboim, Irit Gat-Viks, et al. Digital

cell quantification identifies global immune cell dynamics during influenza infection. *Molecular systems biology*, 10(2):720, 2014.

[3] Rose Du, Vince Carey, and Scott T Weiss. deconvseq: deconvolution of cell mixture distribution in sequencing data. *Bioinformatics*, 35(24):5095–5102, 2019.

[4] Kevin Menden, Mohamed Marouf, Sergio Oller, Anupriya Dalmia, Daniel Sumner Magruder, Karin Kloiber, Peter Heutink, and Stefan Bonn. Deep learning–based cell composition analysis from tissue expression profiles. *Science advances*, 6(30):eaba2619, 2020.

[5] Yating Lin, Haojun Li, Xu Xiao, Lei Zhang, Kejia Wang, Jingbo Zhao, Minshu Wang, Frank Zheng, Minwei Zhang, Wenxian Yang, et al. Daism-dnnxmbd: Highly accurate cell type proportion estimation with in silico data augmentation and deep neural networks. *Patterns*, 3(3):100440, 2022.

[6] Rui Dong and Guo-Cheng Yuan. Spatialdwls: accurate deconvolution of spatial transcriptomic data. *Genome biology*, 22(1):1–10, 2021.

[7] Ying Ma and Xiang Zhou. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature Biotechnology*, pages 1–11, 2022.

[8] Daphne Tsoucas, Rui Dong, Haide Chen, Qian Zhu, Guoji Guo, and Guo-Cheng Yuan. Accurate estimation of cell-type composition from gene expression data. *Nature communications*, 10(1):1–9, 2019.

[9] Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, 10(1):1–9, 2019.

[10] Franziska Görtler, Marian Schoen, Jakob Simeth, Stefan Solbrig, Tilo Wettig, Peter J Oefner, Rainer Spang, and Michael Altenbuchinger. Loss-function learning for digital tissue deconvolution. *Journal of Computational Biology*, 27(3):342–355, 2020.

[11] Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Kristina M Garske, Jae Hoon Sul, Kirsi H Pietiläinen, Päivi Pajukanta, and Eran Halperin. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature communications*, 11(1):1–11, 2020.

[12] Meichen Dong, Aatish Thennavan, Eugene Urrutia, Yun Li, Charles M Perou, Fei Zou, and Yuchao Jiang. Scdc: bulk gene expression deconvolution by multiple single-cell rna sequencing references. *Briefings in bioinformatics*, 22(1):416–427, 2021.

[13] Francisco Avila Cobos, José Alquicira-Hernandez, Joseph E Powell, Pieter Mestdagh, and Katleen De Preter. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature communications*, 11(1):1–14, 2020.

[14] Max Schelker, Sonia Feau, Jinyan Du, Nav Ranu, Edda Klipp, Gavin MacBeath, Birgit Schoeberl, and Andreas Raue. Estimation of immune cell content in tumour tissue using single-cell rna-seq data. *Nature communications*, 8(1):1–12, 2017.

[15] Julien Racle, Kaat de Jonge, Petra Baumgaertner, Daniel E Speiser, and David Gfeller. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *elife*, 6, 2017.

[16] Marian Schön, Jakob Simeth, Paul Heinrich, Franziska Görtler, Stefan Solbrig, Tilo Wettig, Peter J Oefner, Michael Altenbuchinger, and Rainer Spang. Dtd: an r package for digital tissue deconvolution. *Journal of Computational Biology*, 27(3):386–389, 2020.

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[18] Poornima Bhat-Nakshatri, Hongyu Gao, Liu Sheng, Patrick C. McGuire, Xiaoling Xuei, Jun Wan, Yunlong Liu, Sandra K. Althouse, Austyn Colter, George Sandusky, Anna Maria Storniolo, and Harikrishna Nakshatri. A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells. *Cell Reports Medicine*, 2(3):100219, March 2021.

[19] Sunny Z. Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R. Torpy, Nenad Bartonicek, Taopeng Wang, Ludvig Larsson, Dominik Kaczorowski, Neil I. Weisenfeld, Cedric R. Uytingco, Jennifer G. Chew, Zachary W. Bent, Chia-Ling Chan, Vikkitharan Gnanasambandapillai, Charles-Antoine Dutertre, Laurence Gluch, Mun N. Hui, Jane Beith, Andrew Parker, Elizabeth Robbins, Davendra Segara, Caroline Cooper, Cindy Mak, Belinda Chan, Sanjay Warrier, Florent Ginhoux, Ewan Millar, Joseph E. Powell, Stephen R. Williams, X. Shirley Liu, Sandra O'Toole, Elgene Lim, Joakim Lundeberg, Charles M. Perou, and Alexander Swarbrick. A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics*, 53(9):1334–1347, September 2021.

[20] Julien Racle, Kaat de Jonge, Petra Baumgaertner, Daniel E Speiser, and David Gfeller. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife*, 6, November 2017.

[21] José Bragança, Tracey Swingler, Fatima IR Marques, Tania Jones, Jyrki J Eloranta, Helen C Hurst, Toshihiro Shioda, and Shoumo Bhattacharya. Human creb-binding protein/p300-interacting transactivator with ed-rich tail (cited) 4, a new member of the cited family, functions as a co-activator for transcription factor ap-2. *Journal of Biological Chemistry*, 277(10):8559–8565, 2002.

[22] Stephen B. Fox, José Bragança, Helen Turley, Leticia Campo, Cheng Han, Kevin C. Gatter, Shoumo Bhattacharya, and Adrian L. Harris. CITED4 Inhibits Hypoxia-Activated Transcription in Cancer Cells, and Its Cytoplasmic Location in Breast Cancer Is Associated with Elevated Expression of Tumor Cell Hypoxia-Inducible Factor 1α. *Cancer Research*, 64(17):6075–6081, 09 2004.

[23] Jiřina Bártková, Joy Burchell, Jiří Bártek, Bořivoj Vojtěšek, Joyce Taylor-Papadimitriou, Aleš Rejthar, Zdenka Stašková, and Jan Kovařík. Lack of β-casein production by human breast tumours revealed by monoclonal antibodies. *European Journal of Cancer and Clinical Oncology*, 23(10):1557–1563, 1987.

[24] Christen Dillard, Meagan Kiyohara, Vei Mah, Sean P McDermott, Dana Bazzoun, Jessica Tsui, Ann M Chan, Ghassan Haddad, Matteo Pellegrini, Yu-Ling Chang, et al. Emp2 is a novel regulator of stemness in breast cancer cells. *Molecular cancer therapeutics*, 19(8):1682–1695, 2020.

[25] Nicole Salazar and Brian A Zabel. Support of tumor endothelial cells by chemokine receptors. *Frontiers in immunology*, 10:147, 2019.

[26] Renata Mazzucchelli and Scott K Durum. Interleukin-7 receptor expression: intelligent design. *Nature Reviews Immunology*, 7(2):144–154, 2007.

[27] K Muegge, MP Vila, and SK Durum. Interleukin-7: a cofactor for v (d) j rearrangement of the t cell receptor β gene. *Science*, 261(5117):93–95, 1993.

[28] Arthur E Hoerl and Robert W Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

[29] Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

[30] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[31] Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.