Deconvolution of omics data in Python with Deconomix – cellular compositions, cell-type specific gene regulation, and background contributions

Malte Mensching-Buhr[†1], Thomas Sterr[†1], Nicole Seifert[1], Dennis Völkl[2], Jana Tauschke[3], Austin Rayford[4], Helena U. Zacharias[3], Sushma Nagaraja Grellscheid[5], Tim Beissbarth[1], Franziska Görtler[*6], and Michael Altenbuchinger[*1]

[1]Department of Medical Bioinformatics, University Medical Center Göttingen, 37075 Göttingen, Germany
[2]Institute of Theoretical Physics, University of Regensburg, 93053 Regensburg, Germany
[3]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover Medical School, Carl-Neuberg-Str. 1, 30625 Hannover, Germany
[4]Department of Biomedicine and Centre for Cancer Biomarkers, University of Bergen, Thormohlensgt 55, Bergen N-5008, Norway
[5]Computational Biology Unit, Department of Biological Sciences, University of Bergen, Thormohlensgt 55, Bergen N-5008, Norway
[6]Department of Oncology and Medical Physics, Haukeland University Hospital, Haukelandsveien 22, 5021 Bergen, Norway

[†] shared first authors; [*] shared senior authors.
Correspondence: franziska.gortler@uib.no and michael.altenbuchinger@bioinf.med.uni-goettingen.de

Abstract

Summary: Gene expression profiles of heterogeneous bulk samples contain signals from multiple cell populations. Studying variations in their composition can help to identify cell populations relevant for disease. Moreover, analyses, such as the identification of differentially expressed genes, can be confounded by cellular composition, as differences in gene expression may arise from both variations in cellular composition and gene regulation. Here, we present Deconvolution of omics data (Deconomix) – a comprehensive toolbox for the cell-type deconvolution of bulk transcriptomics data. Deconomix stands apart from competing solutions with rich functionality and highly efficient implementations. It facilitates (A) the inference of cellular compositions from bulk transcriptomics data, (B) the machine learning-based optimization of gene weights to resolve small cell populations and to disentangle phenotypically related cells, (C) the inference of background contributions which otherwise would deteriorate cell-type deconvolution, and (D) population estimates of cell-type specific gene regulation.

Availability: Deconomix is available at https://gitlab.gwdg.de/MedBioinf/MedicalDataScience/Deconomix under GPLv3 licensing. The Python package can be easily installed via pip. It comes with a comprehensive documentation of all user-relevant functions and example workflows provided as Jupyter notebooks.

1 Introduction

Heterogeneous bulks are composed of molecularly diverse cellular populations – the cell types. Variations in cell composition in diseases like cancer have been repeatedly observed and their association with clinical or phenotypical data can drive the identification of therapeutic targets (Becker *et al.*, 2022; Wang *et al.*, 2023). Moreover, typical bulk analysis can be confounded by cellular composition (Cook and Vanderhyden, 2019); for instance, in differential gene expression analysis, gene expression differences may arise from both variations in cellular composition and gene regulation. Cell-type deconvolution is a computational strategy to infer cellular compositions from bulk or spatial omics data. Several approaches have been proposed in recent years such as (Altboum *et al.*, 2014; Chen *et al.*, 2018; Menden *et al.*, 2020; Görtler *et al.*, 2024). Moreover, comprehensive benchmark studies were conducted (Avila Cobos *et al.*, 2020; Sutton *et al.*, 2022; Chen *et al.*, 2022).

Numerous issues complicate cell-type deconvolution and those are inextricably linked. First, the inference of small cell populations is challenging. It is well recognized that lowly abundant cell types can be highly relevant in disease. For instance, rare immune cells can be key components in cancer defense. As minor cell populations contribute little to the observed bulk profiles, cell-type deconvolution typically assigns a small cellular proportion to them. However, to establish meaningful associations between cellular abundance and phenotype, class, or patient outcomes, the estimated cellular proportions must also correlate closely with the true proportions. Thus, precise estimates are essential, as coarse estimates can blur association analyses. Second, molecularly related cell types can be hardly disentangled by cell-type deconvolution (Ivich *et al.*, 2024). This difficulty arises when bulk profiles can be reconstructed by different linear combinations of related cell types, which obscures the estimated weights, i.e., the cellular proportions. Third, if specific cell types are missing in the reference profiles, model results are typically compromised (Avila Cobos *et al.*, 2020; Wang *et al.*, 2019). Consequently, all cell types contained in the bulks should ideally be also included in the reference matrix (Avila Cobos *et al.*, 2020). Moreover, if this requirement is not fulfilled, estimated proportions of the remaining cell types will be artificially increased and thus biased (Chu *et al.*, 2022). A forth, related problem is the source of reference profiles. It is widely recognized that cells' molecular phenotypes are influenced by their environment, which also affects results from cell-type deconvolution (Schelker *et al.*, 2017; Racle *et al.*, 2017). To deal with this issue, it was suggested to select references from niches that are comparable to the bulk samples being studied (Schelker *et al.*, 2017). In summary, practical needs motivate optimized solutions for cell-type deconvolution and the inference is challenged by domain differences such as hidden background contributions and environmental factors. Moreover, recent methodological successes have substantially broadened the scope of cell-type deconvolution (Chu *et al.*, 2022; Simeth *et al.*, 2023), and provide estimates of cell-type specific gene regulation from bulk transcriptomics data. One should note that the described obstacles are strongly connected and cannot be solved in isolation; if cell types are missing in the reference profiles, the inference of the remaining, modeled cell-types are biased and, vice versa, incorrect estimates of the modeled cell-types can deteriorate the inference of background contributions.

Recognizing the fact that the aforementioned issues and obstacles are connected, we suggest a comprehensive toolbox – Deconvolution of omics Mixtures (Deconomix) – to address them jointly.

Deconomix is composed of different modules: Module 1 addresses the gene selection and weighting. It uses single-cell data to optimize cell-type deconvolution, addressing also the deconvolution of minor cell populations and molecularly related cell types. Module 2 uses the gene weighting of module 1 to infer both cellular proportions and the cellular background. Finally, Module 3 infers cell-type specific gene regulation from bulk transcriptomics data. Furthermore, each module provides comprehensive model validation and visualization options.

2 Results

2.1 Deconomix analysis modules

We will first briefly introduce the three analysis modules of Deconomix and the main features they comprise.

Module 1 This module is dedicated to the machine-learning based optimization of cell-type deconvolution using single-cell data. Specifically, Deconomix provides different strategies to generate pseudo bulks from single-cell data. These pseudo bulks are artificial mixtures obtained by randomly drawing single cell profiles from the input data and by aggregating their gene expression. Given these mixtures, Deconomix establishes a cell-type deconvolution model for improved estimates of cellular proportions. Here, Deconomix optimizes the Pearson's correlation between estimated and ground truth compositions, as suggested in (Görtler *et al.*, 2020). Moreover, it provides the cosine similarity as alternative optimization strategy. Figure 1A shows the corresponding optimization paths exemplified for Pearson's correlation, where the dotted black line represents the average correlation (y-axis) versus the number of iterations used for model optimization and the colored lines correspond to the cell types included in the reference matrix. Fig 1B shows the corresponding scatter plot for NK cells before (gray) and after optimization (red). The output of this module are gene weights, where genes informative for cell-type deconvolution receive high weights, while those which are irrelevant receive low weights.

Module 2 In this module, the bulk data are analyzed. This Deconomix module requires a gene-weight vector (module 1) and a set of bulk transcriptomic profiles as input. Then, Deconomix performs the cell-type deconvolution, providing both estimates of cellular proportions and background contributions (Fig. 1C and D). The latter consist of a consensus background profile across all bulks and the respective bulk-specific proportions. This module further provides comprehensive functions for model visualization, comprising plot functions to illustrate the cellular proportions and the background profiles.

Module 3 This module is dedicated to the inference of cell-type specific gene regulation. It is well understood that differences in gene expression can be both the consequence of cellular composition and gene regulation. If differential gene expression is the consequence of differential gene regulation, it remains to be elucidated in which cell type a gene is regulated. This is achieved by this module. It requires the bulk profiles to be analyzed, gene weights (module 1), and the cellular proportions and background contributions (module 2) as input. Then, cell-type specific gene regulation can be inferred on a population level using the algorithmic strategy outlined in (Görtler *et al.*, 2024) and in the Methods section. Initially, this module performs a hyper-parameter search to control overfitting (Fig. 1E). Then, the user can select the best hyper-parameter between three different choices for subsequent analysis – one conservative ("lambda.min") and two more liberal choices ("lambda.1se"

and "lambda.max.slope"). The algorithm finally returns a matrix of gene-regulation factors Δ_{jk} for gene j in cell type k (Fig. 1F). For instance, a factor of $\Delta_{jk} = 2$ corresponds to a two-fold increase in gene j in cell type k in the bulks relative to the reference matrix.

2.2 Breast cancer use case

Next, we present a real-world application scenario for Deconomix. We downloaded breast cancer single-cell RNA-Seq data together with an annotation of major and minor cell populations form DISCO (Li *et al.*, 2021). Those were mapped to in total 19 cell populations, and separated into two equally sized sets of single-cell profiles for training and testing ($n = 98453$ vs $n = 98465$) with individual cell populations summarized in supplementary Table S1. Importantly, we removed cancer cells before profile aggregation in the training data, resulting in 67083 cells remaining in the training set, but retained them in the test data to demonstrate Deconomix's capability to infer hidden cell populations. We then generated artificial bulk mixtures by aggregating single-cell profiles (see Methods).

Starting with Module 1, we established gene weights using the training data with results shown in Supplementary Figures S1, where the estimation accuracy in terms of Pearson's correlation is contrasted with the iterations of the optimization procedure for each of the 18 cell types included in the reference matrix. We observed substantial improvements across all populations. For instance, arterial epithelial cells are a minor cell population which constitutes only about 0.51% in the training and 0.35% in the test data of all cells. There, we observed a performance of $r = 0.198$ before optimization and of $r = 0.753$ after optimization. Similarly, the inference of CD4 T-cells and CD8 T-cells, which share a similar molecular profile (Spearman correlation of 0.965), is highly compromised before optimization with r=0.519 and r=0.49, respectively. Deconomix improves results to r=0.882 and r=0.819 after optimization, respectively.

Next, we applied Module 2 on the test data for model verification. Module 2, in contrast to module 1, accounts for potential background contributions. Using the gene weights determined in Module 1, Deconomix yielded the results summarized in Table S2 in terms of Pearson's correlation and Normalized Mean Absolute Error (NMAE). Notably, the test mixtures contain a background contribution attributed to cancer cells, while no cancer cells were included in the training mixtures. In summary, we observed that accounting for a background contributions further improved results, yielding an average performance of $r = 0.736 \pm 0.005$ across all cell population, compared to $r = 0.71 \pm 0.006$ for the Deconomix model without background. Thus, modeling background contributions can improve the inference of the cell types included in the reference matrix. Furthermore, module 2 generates additional information, namely cellular proportions for the background contributions (here cancer cells) and a corresponding consensus background profile. Comparing the inferred background proportions with the respective ground truth using Pearson's correlation yielded $r = 0.553 \pm 0.024$. A comparison between the inferred consensus profile and respective ground truth yielded $r = 0.939 \pm 0.001$. Thus, Deconomix's infers both the molecular background and its contribution to an individual sample.

In previous analysis, we trained and verified Deconomix on pseudo bulks generated from single-cell data. Next, we applied Deconomix on real bulk data from breast cancer patients downloaded from the TCGA Research Network (`https://www.cancer.gov/tcga`). These data comprise four breast cancer subtypes: $n = 190$ triple-negative breast cancers (TNBC; ER−, PR−, HER2−), $n = 82$ human epidermal growth factor receptor 2-positive breast cancer (HER2+), $n = 562$ luminal A (LumA; ER+, PR+, HER2−, and low levels of Ki-67), $n = 209$ luminal B (LumB; ER+, PR+/− and HER2+). Application of Deconomix module 2 yielded cellular proportions for in total 18 cell types included in the reference matrix, as well as background proportions (Suppl. Table S3). In order to

identify subtype specific cellular compositions, we tested for associations between cellular proportions and subtypes using Kruskal-Wallis tests (Suppl. Table S3). The most significant differences were observed for granulocytes ($p < 2 \cdot 10^{-73}$), followed by lymphatic LC ($p < 1 \cdot 10^{-36}$), and NK cells ($p < 1 \cdot 10^{-35}$). A direct inspection of the results shows, e.g., that aggressive subtypes like TNBC and HER2+ present with a tumor microenvironment with on average higher infiltration of macrophages, dendritic cells and plasma cells, compared to the less aggressive subtypes. The least aggressive subtype luminal A showed the lowest proportions of tumor epithelial cells and highest proportions of smooth muscle cells, at average (see also Suppl. Table S2).

Finally, we demonstrate Deconomix module 3 and derive cell-type specific gene regulation in TCGA breast cancer bulk. In this study, we retrieved single-cell data of healthy breast tissue from (Li *et al.*, 2021) to establish a deconvolution model via Deconomix module 1, as previously. Eight major cell populations were included in the reference matrix: B cells, CD4+ T cells, CD8+ T cells, endothelial cells, fibroblasts, myeloid cells, NK cells, and perivascular cells. Those references represent cells in a cancer free environment. Thus, our analysis infers cell-type specific gene regulation relative to healthy controls. The TCGA breast cancer dataset comprises a total of 4,851 genes. To ensure consistency between the single cell and bulk data, the overlap of genes present in both sets was determined, resulting in 957 genes. Additionally, MALAT1 was excluded from the gene pool due to its expression level being four (single cell) to ten (bulk) times higher than that of the second most highly expressed gene. Single highly expressed genes could introduce a significant bias in the downstream analysis, potentially overshadowing biologically relevant effects. To begin with, we verified model performance as previously (Suppl. Table S6 and Fig. S3.). Subsequently, we applied Module 3 on the TCGA breast cancer data, where a comprehensive hyper-parameter search is performed to control overfitting (see Methods, Supp. Fig. S4). Given the selected hyper-parameter, Module 3 returns cell-type specific gene regulation for the four breast cancer subtypes. First, we explored the top up-regulated genes and their overlap. Among the top 50 genes, 29 genes were affected in all subtypes. Considering the top 20 genes, only, the overlap reduced to eight genes: B2M, IGHA2, IGHA1, IGHD, JCHAIN, JUNB, FCER1A and JUND, as further illustrated in Fig. 2. In addition, 15 genes were affected in at least 3 of the four subtypes. Deconomix further provides comprehensive functionality for downstream analysis, comprising a pipeline for enrichment analysis. As such, Deconomix can test, whether specific gene groups are particularly regulated in a given cell type. For illustration purposes, we performed such an analysis using two pre-specified gene groups: "CCL" (chemokine ligands) and "GZM" (granzymes) genes. We observed that both gene groups in all four breast cancer subtypes were significantly up-regulated in CD8+ T cells (p-values below 0.01). Interestingly, granzyme genes encode enzymes active in cytotoxic T cells (CD8+ T cells) and as such play a critical role in cancer defense Zöphel *et al.* (2022). Further, in myeloid cells, CCL genes were significantly up-regulated in all subtypes. When myeloid cells infiltrate the tumor, they produce CCL genes such as Chemokine (C-C motif) ligands 2 (CCL2) Kohli *et al.* (2022). A heatmap illustrating the top 20 up-regulated genes is shown in Figure 3. Interestingly, all subtypes show upregulation of several immunoglobulin genes in B cells. IGHG1, for example, is already known to promote angiogenesis, proliferation and invasion of breast cancer cells Larsson *et al.* (2020); Zhang *et al.* (2023).

3 Discussion

Computational approaches for cell-type deconvolution have evolved substantially in recent years, starting from naive strategies based on ordinary least squares to complex models which infer not just cellular proportions, but also potential background contributions and cell-type specific gene

regulation (Chu *et al.*, 2022; Görtler *et al.*, 2024; Simeth *et al.*, 2023). However, the implementation of such complex models requires dedicated algorithmic solutions. Here, we combined our theoretical and algorithmic insights from (Görtler *et al.*, 2020; Görtler *et al.*, 2024), extended them by rich functionality and provide them as a user friendly python package called Deconomix. Deconomix has little requirements to the user and to the computing resources (module 1 to 3). Additionally, we provide expert features for users with comprehensive computing resources. Moreover, the Deconomix package comes with implementations to generate pseudo bulks from single-cell data, to visualize and benchmark cell-type deconvolution, and to perform gene-set enrichment analysis exploring cell-type specific gene regulation.

In summary, Deconomix is a comprehensive toolbox for cell-type deconvolution, providing state-of-the-art algorithmic solutions for the inference of cellular proportions, background contributions, and cell-type specific gene regulation.

4 Methods

4.1 Nomenclature

Let $\boldsymbol{y}_i = (y_{1i}, \ldots, y_{pi})^T \in \mathbb{R}_+^{p\times 1}$ represent the bulk RNA profile of sample i containing gene expression levels for p genes on a natural scale (not on log scale). Further, let $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \in \mathbb{R}_+^{p\times n}$ be a matrix containing n bulk profiles in its columns. The archetypical cell-type reference profiles are denoted as $\boldsymbol{x}_k = (x_{1k}, \ldots, x_{pk})^T \in \mathbb{R}_+^{p\times 1}$ with $k = 1, \ldots, q$ and $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q) \in \mathbb{R}_+^{p\times q}$ is the reference matrix containing the q reference profiles in its columns. We further use the nomenclature $\mathbf{A}_{k,\cdot}$ to indicate the kth row of an arbitrary matrix $\mathbf{A}$.

4.2 Loss-function learning for cell-type deconvolution

The key idea behind cell-type deconvolution is to express the observed bulk profiles as a linear combination of reference profiles, $\boldsymbol{y}_i = \mathbf{X}\boldsymbol{c}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{c}_i = (c_{i1}, \ldots, c_{iq})^T$ are the regression weights for the q cell types in sample i and $\boldsymbol{\epsilon}_i = (\epsilon_{1i}, \ldots, \epsilon_{pi})$ the corresponding gene residuals. For a set of bulk profiles, this equation becomes

$$\mathbf{Y} = \mathbf{X}\mathbf{C} + \mathbf{E}\,, \tag{1}$$

where $\mathbf{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n) \in \mathbb{R}_+^{q\times n}$ contains the regression weights for n samples in its columns. The corresponding model residuals are collected in $\mathbf{E} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n) \in \mathbb{R}^{p\times n}$. The regression weights can be interpreted as cellular compositions under normalization constraints; if each column of $\mathbf{Y}$ and $\mathbf{X}$ sums up one and assuming that there are no missing reference profiles in $\mathbf{X}$, one can directly derive that $\sum_k c_{ik} = 1$ for $i = 1, \ldots, n$ (Racle *et al.*, 2017; Görtler *et al.*, 2024). For computational efficiency, we relied on non-negative least squares,

$$\hat{\mathbf{C}} = \arg\min_{\mathbf{C}\succeq 0} \mathcal{L}(\mathbf{C}) = \frac{1}{n}\left\|\mathbf{Y} - \mathbf{X}\mathbf{C}\right\|_F^2\,, \tag{2}$$

which can also be minimized for each $\boldsymbol{c}_i$ individually. The key step of loss-function learning is to generalize the loss to

$$\mathcal{L}(\mathbf{C}, \Gamma) = \frac{1}{n}||G(\mathbf{Y} - \mathbf{X}\mathbf{C})||_F^2\,, \tag{3}$$

with $G = \mathrm{diag}(g_1, \ldots, g_p) \in \mathbb{R}^{p\times p}$ containing weights g_j for each gene j. Then, for each gene j, the corresponding term in the sum of squared residuals is weighted by a factor $\gamma_j = g_j^2$. Thus, to improve estimates of cellular compositions, genes can be weighted according to their importance for cell-type deconvolution; genes which improve cell-type deconvolution are up-weighted while those which don't are down-weighted. This is achieved by optimizing an outer loss function, for which we

provide two different choices. In line with (Görtler *et al.*, 2020), we implemented the outer objective function

$$L(\Gamma) = \frac{1}{q} \sum_{k=1}^{q} \mathrm{cor}(\mathbf{C}_{k,\cdot}, \hat{\mathbf{C}}_{k,\cdot}(\Gamma)) \,, \tag{4}$$

which is maximized with respect to $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_p)$ subject to $\sum_{j=1}^{p} \gamma_j = 1$, given a reference matrix $\mathbf{X}$ and training mixtures $\mathbf{Y}$ of known cellular composition $\mathbf{C}$. To accelerate optimization, $\hat{\mathbf{C}}$ is determined from ordinary least squares, where all values $c_{ik} \leq 0$ are set to zero. The training mixtures can be generated, e.g., using single-cell data (Görtler *et al.*, 2020).

One should note that $\mathrm{cor}(\mathbf{C}_{k,\cdot}, \hat{\mathbf{C}}_{k,\cdot}(\Gamma))$ is the correlation between ground truth proportions and their estimates for cell type k. Consequently, each modeled cell type is equally important for the outer loss, which is crucial for the correct inference of minor and molecularly related cell populations. As an alternative choice for the outer loss function, we implemented the cosine similarity given by

$$L(\Gamma) = \frac{1}{q} \sum_{k=1}^{q} \frac{\mathbf{C}_{k,\cdot} \cdot \hat{\mathbf{C}}_{k,\cdot}(\Gamma)}{\|\mathbf{C}_{k,\cdot}\|_2 \|\, \hat{\mathbf{C}}_{k,\cdot}(\Gamma)\|_2} \,. \tag{5}$$

For implementation of both outer loss functions, the Pytorch machine learning library was used (Paszke *et al.*, 2019). Respective functions for gene-weight optimization are provided in Deconomix module 1.

4.3 Adaptive Digital Tissue Deconvolution (ADTD)

Adaptive Digital Tissue Deconvolution (ADTD) built on loss function learning (Görtler *et al.*, 2024). It uses the gene weights Γ determined as outlined, but additionally accounts for potential background contributions not contained in the reference matrix and for environmental factors. The ADTD loss is defined as

$$\begin{aligned} \mathcal{L}_{\mathrm{ADTD}}(\mathbf{C}, \tilde{\mathbf{c}}, \tilde{\mathbf{x}}, \Delta) = \|G(\mathbf{Y} - (\Delta \circ \mathbf{X})\mathbf{C} - \tilde{\mathbf{x}}\tilde{\mathbf{c}})\|_F^2 \\ + \lambda_1 \|\mathbf{C} - \mathbf{C}_0\|_F^2 + \lambda_2 \|J_{p,q} - \Delta\|_F^2 \,. \end{aligned} \tag{6}$$

The corresponding optimization problem is

$$\hat{\mathbf{C}}, \hat{\tilde{c}}, \hat{\tilde{x}}, \hat{\Delta} = \underset{\mathbf{C} \succeq 0_{q,n}, \tilde{\mathbf{c}} \succeq 0_{1,n}, \tilde{\mathbf{x}} \succeq 0_{p,1}, \Delta \succeq 0_{p,q}}{\arg\min} \{\mathcal{L}_{\mathrm{ADTD}}(\mathbf{C}, \tilde{\mathbf{c}}, \tilde{\mathbf{x}}, \Delta)\} \,,$$

subject to $\sum_{j=1}^{p} \tilde{\mathbf{x}}_j = 1$, $\sum_{j=1}^{p} (\Delta \circ \mathbf{X})_{jk} = 1$ for all k, and $\tilde{\mathbf{c}}_i + \sum_{k=1}^{q} \mathbf{C}_{ki} = 1$. Here, the term $\tilde{\mathbf{x}}\tilde{\mathbf{c}} \in \mathbb{R}_+^{p \times n}$ models background contributions, where the unknown background $\tilde{\mathbf{x}} \in \mathbb{R}_+^{p \times 1}$ contributes with proportions $\tilde{\mathbf{c}} \in \mathbb{R}_+^{1 \times n}$ for the n bulks. Further, $\Delta \circ \mathbf{X}$ represents an adapted reference matrix, where $\Delta \in \mathbb{R}_+^{p \times q}$ contains gene and cell-type specific re-scaling factors which are multiplied element wise with the entries of $\mathbf{X}$ via the Hadamard product $\circ$. The re-scaling factors Δ_{jk} can be interpreted as cell-type specific gene regulation; a coefficient of $\Delta_{jk} = 2$ corresponds to a two-fold increase, $\Delta_{jk} = 1$ to no gene regulation, and $\Delta_{jk} = 0.5$ to a halved expression in gene j and cell type k relative to the reference, respectively. Eq. (6) further contains two regularization terms. The term $\lambda_1 \|\mathbf{C} - \mathbf{C}_0\|_F^2$ regularizes the cellular proportions $\mathbf{C}$ and biases their estimates to pre-specified proportions $\mathbf{C}_0$. DeconomiX determines $\mathbf{C}_0$ by minimizing the ADTD loss with $\Delta_{jk} = 1$ for all j, k. Thus $\mathbf{C}_0$ corresponds to the ADTD solution without environmental adaptation. The solution $\hat{\mathbf{C}} = \mathbf{C}_0$ is retrieved for $\lambda_1 \to \infty$. The second regularization term $\lambda_2 \|J_{p,q} - \Delta\|_F^2$ biases the environmental adaptation factors towards the solution without adaptation ($\Delta_{jk} = 1$ for all j, k).

4.4 Deconomix hyper-parameter tuning

Deconomix module 2 assumes that the reference profiles do not require adaptation ($\Delta_{jk} = 1$ for all j, k). Thus, the user directly obtains the ADTD solution for $\lambda_1 \to \infty, \lambda_2 \to \infty$. Despite its simplicity, this solution already provides a remarkably accurate estimate of cellular proportions $\mathbf{C}$, the background profile $\tilde{\mathbf{x}}$ and respective background proportions $\tilde{\mathbf{c}}$ (Görtler *et al.*, 2024).

Module 3 is intended to infer population estimates of cell-type specific gene regulation. Those are encoded in the adaptation factors Δ. In order to provide a computationally cheap strategy to infer Δ without extensive hyper-parameter screening, we used the following strategy:

- Set $\mathbf{C}$ to the solution without cellular adaptation ($\mathbf{C} = \mathbf{C}_0$).
- Set the gene weights to

$$G = \frac{1}{\sqrt{n}} \operatorname{diag}(1/\text{mean}(Y_{1,\cdot})^2, \ldots, 1/\text{mean}(Y_{p,\cdot})^2)\,.$$

  This choice ensures that genes contribute comparable to the optimization problem and that a single λ_2 value can be used for different classes.
- Split the data into batches. When Deconomix is used with standard settings, the data are split in five batches of approximately equal size.
- Use $k-1$ of the k data batches to learn an ADTD model for a series of λ_2 values and record the loss on the left-out batch. Repeat this procedure such that every batch served for model validation once (k-fold cross validation).

Former procedure gives an estimate of the generalization error together with error intervals. Deconomix provides three choices to select λ_2. The first choice is named "lambda2.min" and returns the λ_2 value with the lowest generalization error. The second choice is more conservative and is named "lambda2.1se". It is defined as the maximum value that λ_2 can take such that it is still within the one standard error interval obtained for $\lambda_2 =$ lambda.min, in line with (Friedman *et al.*, 2010). The third choice 'lambda.max.slope" is also more conservative than "lambda.min" and is defined by the maximal slope in λ_2-log space within the interval between "lambda.min" and the maximal considered value. The latter choice was used for the presented use case analysis. Further, if multiple data sets were analyzed simultaneously, we selected the most conservative hyper-parameter determined across data sets.

Module 4 performs comprehensive hyper-parameter screening using a grid of λ_1 and λ_2 values. The grid can be pre-specified by the user and in contrast to module 3, the procedure requires test data with known cellular compositions. Thus, this procedure is only feasible if respective data are available for the target domain (e.g., from single-cell RNA-Seq).

4.5 Breast cancer use case

Data simulation We generated $20,000$ artificial bulk mixtures for training by randomly drawing 100 cell profiles each from the training cohort (Table S1). Further, we generated 10 batches for model testing with $1,000$ samples each, following the same strategy. We then aggregated the training single-cell profiles to consensus cell-type profiles for each cell population and calculated gene variances across the reference profiles. Finally, we retained the $5,000$ genes with highest variance for analyses and restricted our reference matrix, and the training and test mixtures to those genes.

Domain transfer For the domain transfer example, we generated $5,000$ artificial bulk mixtures by drawing from the training cohort (Table S5) and 10 times $1,000$ bulk mixtures by drawing from the test cohort. The drawing was like for the use case described in Data simulation. As we had only 8 immune cell types, we regarded here only the $1,000$ genes with the highest variance.

5 Competing interests

None declared.

6 Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (grant numbers: 01ZX1912A and 01ZX1912C) and by the DFG SFB-TRR 274.

References

Altboum, Z. *et al.* (2014). Digital cell quantification identifies global immune cell dynamics during influenza infection. *Molecular systems biology*, **10**(2), 720.

Avila Cobos, F. *et al.* (2020). Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature communications*, **11**(1), 1–14.

Becker, W. R. *et al.* (2022). Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. *Nature Genetics*, **54**(7), 985–995.

Chen, B. *et al.* (2018). Profiling Tumor Infiltrating Immune Cells with CIBERSORT. In *Cancer systems biology*, pages 243–259. Springer.

Chen, J. *et al.* (2022). A comprehensive comparison on cell-type composition inference for spatial transcriptomics data. *Briefings in Bioinformatics*, **23**(4), bbac245.

Chu, T. *et al.* (2022). Cell type and gene expression deconvolution with bayesprism enables bayesian integrative analysis across bulk and single-cell rna sequencing in oncology. *Nature Cancer*, **3**(4), 505–517.

Cook, D. P. and Vanderhyden, B. C. (2019). Ovarian cancer and the evolution of subtype classifications using transcriptional profiling†. *Biology of Reproduction*, **101**(3), 645–658.

Friedman, J. *et al.* (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1), 1–22.

Görtler, F. *et al.* (2020). Loss-Function Learning for Digital Tissue Deconvolution. *Journal of Computational Biology*, **27**(3), 342–355.

Görtler, F. *et al.* (2024). Adaptive digital tissue deconvolution. *Bioinformatics*, **40**(Supplement_1), i100–i109.

Ivich, A. *et al.* (2024). Missing cell types in single-cell references impact deconvolution of bulk data but are detectable. *bioRxiv*.

Kohli, K. *et al.* (2022). Key chemokines direct migration of immune cells in solid tumors. *Cancer Gene Therapy*, **29**(1), 10–21.

Larsson, C. *et al.* (2020). Prognostic implications of the expression levels of different immunoglobulin heavy chain-encoding rnas in early breast cancer. *NPJ breast cancer*, **6**, 28.

Li, M. *et al.* (2021). Disco: a database of deeply integrated human single-cell omics data. *Nucleic Acids Research*, **50**(D1), D596–D602.

Menden, K. *et al.* (2020). Deep learning–based cell composition analysis from tissue expression profiles. *Science Advances*, **6**(30).

Paszke, A. *et al.* (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Racle, J. *et al.* (2017). Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife*, **6**.

Schelker, M. *et al.* (2017). Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nature communications*, **8**(1), 1–12.

Simeth, J. *et al.* (2023). Virtual tissue expression analysis. *bioRxiv*.

Sutton, G. J. *et al.* (2022). Comprehensive evaluation of deconvolution methods for human brain gene expression. *Nature Communications*, **13**(1), 1358.

Wang, X. *et al.* (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, **10**(1), 1–9.

Wang, Y. *et al.* (2023). Multidirectional characterization of cellular composition and spatial architecture in human multiple primary lung cancers. *Cell Death & Disease*, **14**(7), 462.

Zhang, Y. *et al.* (2023). Ighg1 promotes malignant progression in breast cancer cells through the regulation of akt and vegf signaling. *Biomolecules & biomedicine*, **23**(4), 616–623.

Zöphel, D. *et al.* (2022). Faster cytotoxicity with age: Increased perforin and granzyme levels in cytotoxic cd8+ t cells boost cancer cell elimination. *Aging Cell*, **21**(8), e13668. e13668 ACE-21-0655.R1.

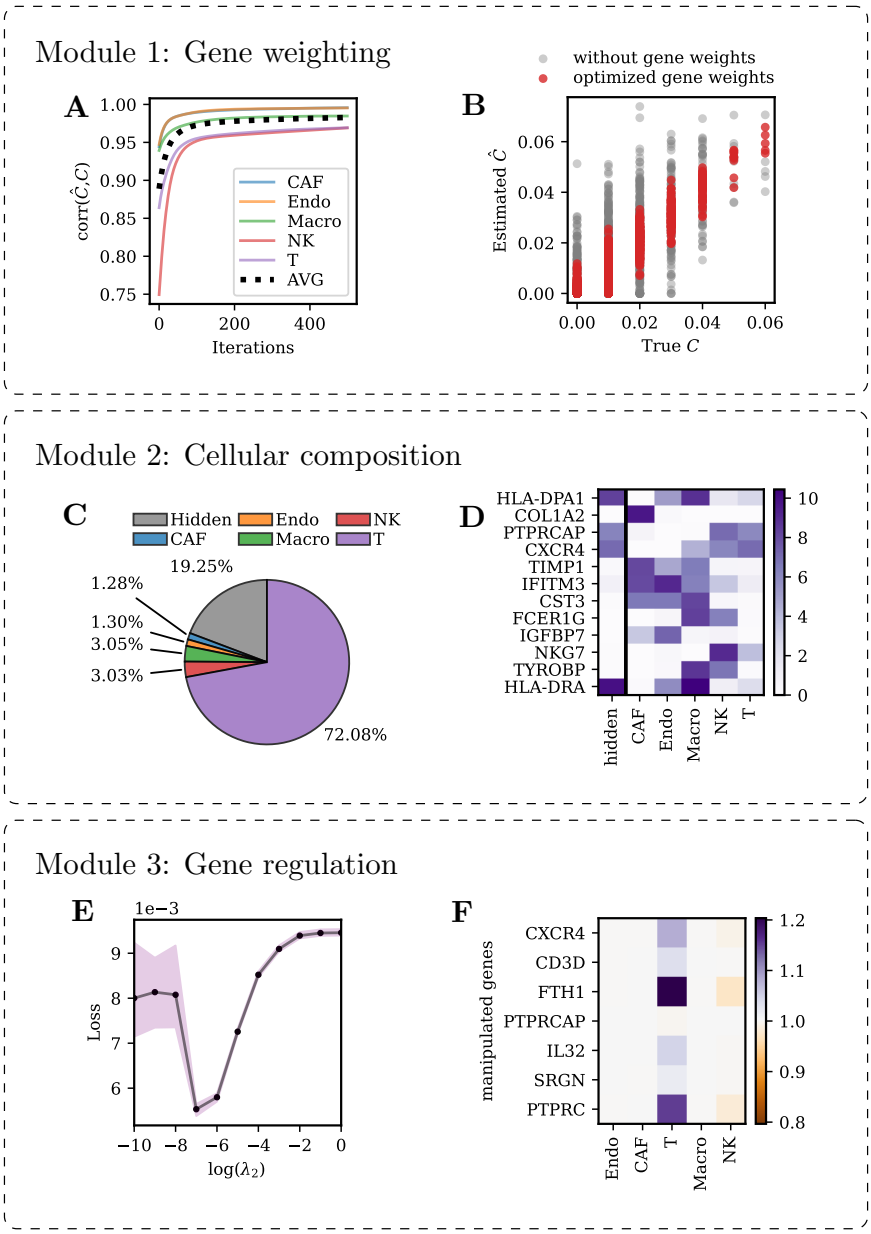

Figure 1: Deconomix overview, including example realizations for each module: (A) Average and cell-type specific correlation between inferred cellular proportions and respective ground-truth values (y-axis) versus the iteration of the Deconomix optimization procedure (x-axis). (B) Scatter plot contrasting predictions (y-axis) with ground truth values (x-axis) before (black) and after Deconomix optimization (red) for a rare cell population (NK cells). (C) Example realization of an inferred cell composition including background contributions. (D) Excerpt of the corresponding reference matrix including an inferred background profiles (column "hidden"). (E) Hyper-parameter grid search to infer cell-type specific gene regulation and the corresponding, and (F) the corresponding cell-type specific gene regulation matrix.

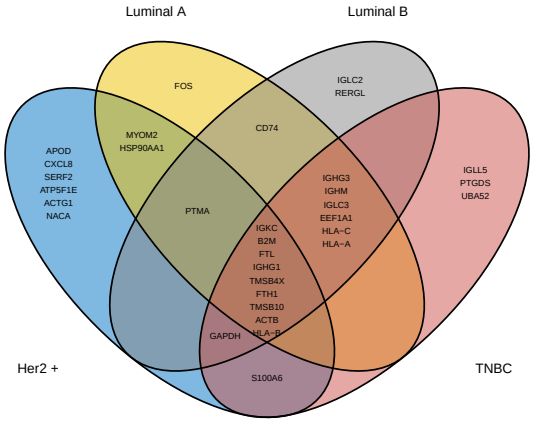

Figure 2: Venn diagram illustrating the overlap of the 20 top up-regulated genes in any cell type across the four breast cancer subtypes.

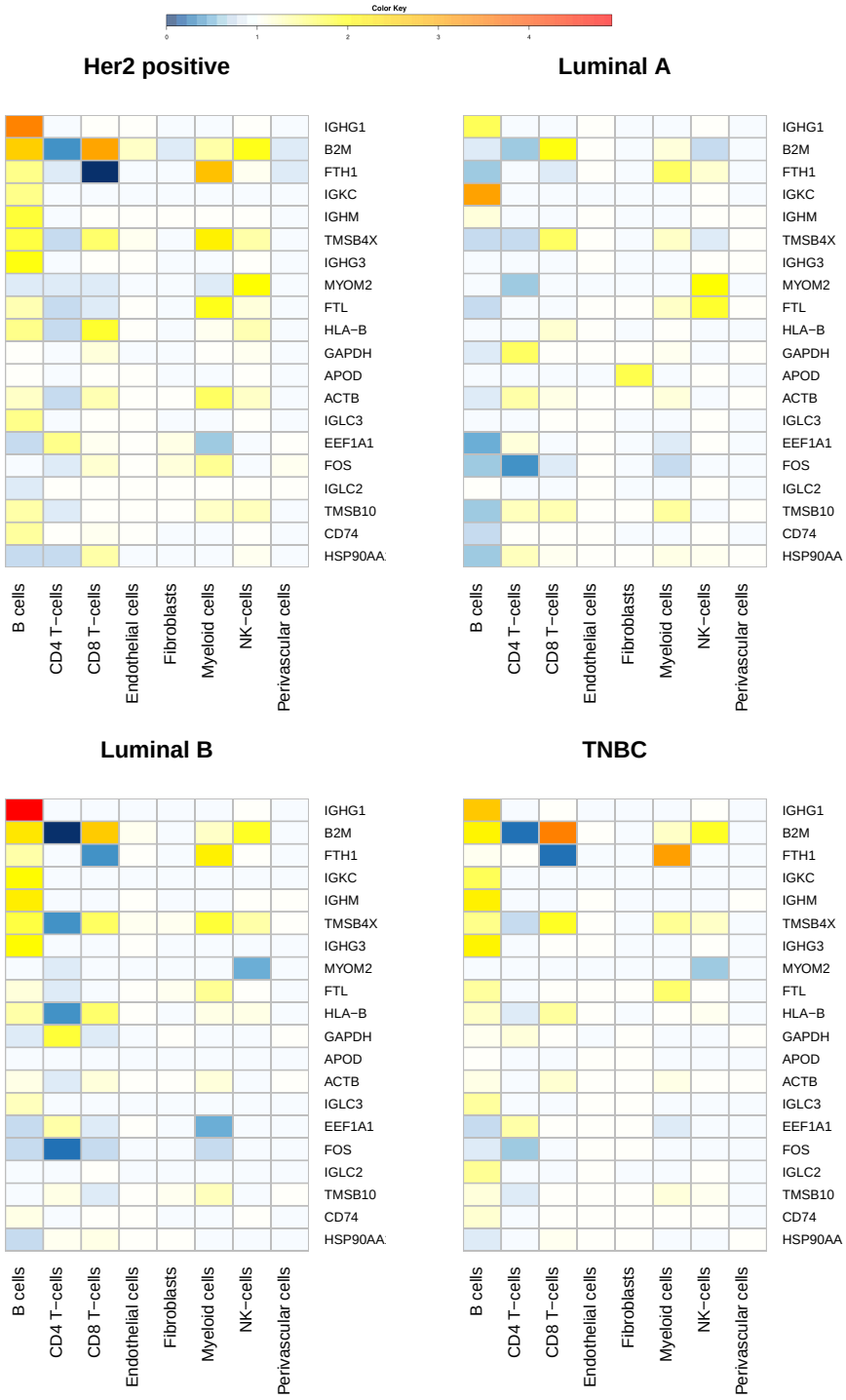

Figure 3: **Heatmap of the 20 genes most upregulated across the four beast cancer subtypes.** Values between 0 and 1 (blue) correspond to down-regulation, white to no regulation, values above 1 (yellow over orange and red) correspond to positive gene regulation.