Preprint, 2026, 1–18

doi: xx.xxxx/xxxx
Manuscript in Preparation
Tech Note

TECH NOTE

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

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

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

[†]Contributed equally.
[*]Shared senior authors. Correspondence: franziska.gortler@uib.no and michael.altenbuchinger@bioinf.med.uni-goettingen.de

Abstract

Background: Gene expression profiles derived from heterogeneous bulk samples contain signals from various cell populations. Cell-type deconvolution approaches are computational tools to reverse engineer the composition of bulks in term of cell populations. Accurate estimates of cell compositions are crucial for identifying 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.
Results: We present Deconvolution of omics data (Deconomix) – a comprehensive toolbox for the cell-type deconvolution of bulk transcriptomics data, available as a Python package and standalone graphical user interface. 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. To showcase the application of Deconomix, we present a case study on breast cancer data from TCGA, highlighting subtype-specific cellular compositions and cell-type-specific gene-regulatory programs.
Conclusion: We present Deconomix, a comprehensive Python package including a graphical user interface for the inference of cellular compositions, cell-type-specific gene regulation, and background contributions from bulk transcriptomics data.

Key words: Cell-type deconvolution, Bulk transcriptomics, Gene regulation, Gene expression analysis, Machine learning, Cellular composition inference

Key Points

- Deconomix optimizes gene weights to disentangle small cell populations and phenotypically related cells.
- Deconomix estimates cell compositions, unknown background contributions and cell-type specific gene regulation from bulk transcriptomics data.
- A Python package is available as code repository and installable via PyPI [1].
- A standalone graphical user interface is available as code repository [2].
- An exemplary analysis of a breast cancer case study is provided as tutorial [3].

Introduction

Heterogeneous bulk transcriptomics samples are composed of molecularly diverse cellular populations – the different cell types. Variations in cell type composition in diseases like cancer have been repeatedly observed and their association with clinical or phenotypical data can drive the identification of therapeutic targets [4, 5]. Moreover, typical bulk analysis can be confounded by cellular composition [6]; for instance, in differential gene expression analysis, gene expression differences may arise from variations in both 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 BayesPrism, CIBERSORTx, Scaden, and ADTD [7, 8, 9, 10]. Moreover, comprehensive benchmark studies were conducted, highlighting performance and functional differences between methods [11, 12, 13].
Numerous issues complicate cell-type deconvolution and those are inextricably linked. First, the inference of small cell populations is challenging. However, lowly abundant cell types can be highly relevant to answer disease-related questions. For instance, rare immune cells can be key components in cancer defense [14]. 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 [15]. This difficulty arises when bulk profiles can be equally well reconstructed by different linear combinations of related cell types. As a consequence, the estimated weights corresponding to cellular proportions become imprecise. Third, if specific cell types are missing in the reference profiles, model results can be compromised [11, 16]. Consequently, all cell types contained in the bulk samples should ideally be also included in the reference matrix [11]. Moreover, if this requirement is not fulfilled, estimated proportions of the remaining cell types will be artificially increased and thus biased [7]. A fourth, related problem is the experimental 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 [17, 18]. To deal with this issue, it was suggested to select references from niches that are comparable to the bulk samples being studied [17]. 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 [7, 19], 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 implements a broad collection of functions that provide methods and utilities for constructing complete cell-type deconvolution pipelines in Python. Its functionality can be organized into three modules: module 1 addresses the gene selection and weighting. Using single-cell training data, it optimizes a set of gene weights on artificially generated bulk profiles and establishes a reference expression matrix for the cell types present in the training data. This procedure allows to reliably quantify lowly abundant and molecularly related cell types. Module 2 uses these gene weights to infer both cellular proportions and the cellular background from bulk transcriptomics data. Module 3 extends the latter analysis by inferring cell-type specific gene regulation. Moreover, Deconomix provides rich functionality to assess and visualize deconvolution results.

Findings

Deconomix analysis modules

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

Gene weight optimization. This module is dedicated to the machine-learning based optimization of cell-type deconvolution using single-cell data. Specifically, Deconomix provides a method to generate pseudo bulks from single-cell data, accompanied by a reference expression matrix for all cell types present in the dataset. The pseudo bulks are artificial mixtures obtained by randomly drawing single cell profiles from the input data and by aggregating their gene expression. Given the reference matrix and the mixtures, Deconomix establishes an optimized cell-type deconvolution via a gene weighting: genes which are informative for cell-type deconvolution receive high weights, while those which are irrelevant receive low weights. To achieve this, Deconomix optimizes the Pearson's correlation between estimated and ground truth compositions, as suggested in [20]. Further, the cosine similarity as alternative optimization strategy. Figure 1A shows the corresponding optimization paths exemplified

Compiled on: March 18, 2026.
Draft manuscript prepared by the author.

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. Figure 1B shows the corresponding scatter plot for a rare cell population before (gray) and after optimization (red).

Cellular composition. In this module, bulk data are analyzed utilizing the optimized gene weights and the reference matrix from the first module. It performs both cell-type deconvolution to estimate cellular proportions and infers potential hidden background contributions (referred to as Deconomix+h, Figure 1C and 1D). Estimates of background contributions consist of a consensus background profile across all bulks and the respective bulk-specific proportions. These estimates are optimized via quadratic programming, as proposed in [10] and described in the methods section. The analysis functions are accompanied by functions for model visualization and validation.

Cell-type-specific gene regulation. This module is designed to infer cell-type-specific gene regulation along potential background contributions (referred to as Deconomix+h,r). Specifically, it disentangles whether differences in gene expression result from variations in cellular compositions or gene regulation. It requires gene weights and reference profiles established in the first module, and bulk profiles to be analyzed as input. Then, cell-type specific gene regulation is estimated relative to the reference profiles. The algorithmic strategy is outlined in [10] and in the Methods section. This module further performs a hyperparameter search to control overfitting (Figure 1E). The user can select the best hyper-parameter between two different choices for subsequent analysis – a liberal choice, the minimum of the curve, and a more conservative choice, the one standard error rule, described in the methods section. Finally, the Deconomix+h,r model, applied with the chosen hyperparameter, returns the cellular compositions for known cell types and the hidden background with the hidden consensus profile. A matrix of gene-regulation factors Δ_{jk} for gene j in cell type k (Figure 1F) is returned, which can be related to cell-type specific gene expression changes: a factor $\Delta_{jk} > 1$ indicates an expression increase in gene j in cell type k in the bulks relative to the reference profile for cell type k.

Deconomix graphical user interface

Besides the Python package, Deconomix comes with its own graphical user interface (GUI) version. In addition to standalone software for Linux and MacOS, we provide the source code [2] to enable hosting the application on a server. The GUI is designed for non-expert users who want to apply our models to their data without requiring programming expertise.

Workflow. Figure 2 shows parts of the GUI and their functionality. The first step is the data import via the input mask shown in Figure 2A. This can be done either via the widely used anndata format (.h5ad) or one can use the file converter tool to generate (.dcx) files, an efficient file format developed for the Deconomix GUI. After importing reference profiles, training and test data, one can start the analysis by optimizing gene weights. In the interface shown in Figure 2C, the user can specify all adjustable parameters, such as number of iterations, number of cells in the artificial bulk mixtures, and number of mixtures. The application then produces pie charts with the cell type distribution for every mixture inferred by a simple Deconomix model, and an interactive heatmap of the top n gene weights, where n can be specified by the user in the plot panel. A third plot provides information about the correlation between estimated cell type distribution and the simulated ground truth via a cell-type-specific scatter plot. Utilizing these learned gene weights, one can establish more complex Deconomix models, resolving possible hidden cell populations and gene regulation patterns. The Deconomix and the Deconomix+h model can be applied directly. For the Deconomix+h,r model, a hyperparameter can be specified (Figure 2D), or the built-in hyperparameter search can be used. After running the models, the software provides visualizations for the results, including a pie chart for the cellular composition of each analyzed bulk sample. For the cell-type specific gene regulation factors inferred with the Deconomix+h,r model, an interactive heatmap is generated that allows visualization for a user-specified gene set. All plots and results can be exported and used for downstream analyses.

Session Manager. Upon starting the GUI, the user can either choose to remain in the automatically created session or decide to start a separate working environment via the built-in session manager, shown in Figure 2B. The session manager is especially useful when multiple experiments need to be analyzed in a structured way. But its full potential is shown, when the Deconomix GUI is hosted on a server. The GUI is then running on a specified port (default 8050) and can be accessed through any browser in the local system. The application can then be used by multiple people, administrated by the session manager. This has several advantages. The GUI version can be hosted on a high performance computing cluster to benefit from increased computational resources in terms of RAM and CPU cores. Further, the GUI can be hosted in a secure server environment to handle sensitive data. Then, the data flow is limited to the local network. An additional benefit is the ease of maintenance, as soon as updates for Deconomix are available, only one instance of the GUI version needs to be updated.

Breast cancer use case

Next, we demonstrate the capabilities of Deconomix in a breast cancer case study, exploring three different settings. The first two settings use single-cell data of breast cancer specimens downloaded from the DISCO database [21, 22]. Those were used to establish artificial bulks for both model training and validation. Both scenarios have the advantage that a clear ground truth is available and that models can be validated accordingly. In contrast, scenario three addresses a real world application on breast cancer bulks from TCGA ([23, 24]), where the single-cell data from DISCO are used for model establishment only. In this scenario, no rigorous validation is feasible, since no established ground truth is available. However, several biological plausibility checks will be provided.

Furthermore, one should note that Deconomix is a comprehensive collection of cell-type deconvolution methods benchmarked previously [10, 20], with modules addressing gene selection and weighting, background inference and the estimation of cell-type specific gene expression. Thus, it contains components that can be used independently, jointly, or even in combination with alternative methodology. For this reason, we refer to the original publications for detailed benchmark analyses, and focus on a typical workflow comprising model development, evaluation, and application to a real world scenario. The Deconomix models used in this study are listed in Table 1, together with their algorithmic background and the quantities they can estimate. All relevant data resources and the necessary preprocessing steps are summarized in the methods section.

Scenario 1: A controlled simulation. Scenario 1 is a controlled setting, where we retrieved only cancer cases from DISCO together with corresponding cell-type annotations aggregated into minor cell populations (Table S1). We first split the data randomly into a training and a test set with compositions summarized in Table S2. For illustration purposes, we removed all samples labeled as cancer epithelial cells from the training set, while we retained them in the test data to simulate a domain shift in terms of an unknown expression contribution. Using Deconomix functions, we generated 20,000 artificial bulk mixtures on the training set and determined an average expression profile for each cell type contained in the training data to establish a reference matrix. For validation, we generated 10,000 artificial bulks from the test data and split them into ten test folds for validation. Then, we applied Deconomix to (1) perform a naive cell type deconvolution with equally weighted genes and the afore derived reference matrix, (2) optimize gene weights using the artificial training bulks, and (3) infer a potential background contribution (Deconomix+h). Scenario 1 offers a clear ground truth in terms of underlying cellular compositions in the test bulks. Therefore, we evaluated models (1) to (3) on the test data and calculated Spearman's correlation coefficient ρ between estimated and true cellular proportions with results summarized in Table S3. The results nicely illustrate the performance benefit of each individual component: while the naive model yielded an average ρ of 0.335 across 18 known cell types, Deconomix with optimized gene weights significantly improved this to an average ρ of 0.634. Notably, B cells improved from 0.185 to 0.793. The naive model further struggled with CD4+ and CD8+ T-cells, yielding 0.414 and 0.319, while Deconomix boosted these values to 0.813 and 0.685. We further observed that Deconomix+h further enhanced performance, achieving an average ρ of 0.658 for known cell types and reliably estimating cellular proportions of an unknown cell type with $\rho = 0.655$. Thus, in summary, we observed that gene-weight optimization improves deconvolution results and that results become more reliable if background contributions are taken into account.

Scenario 2: A controlled simulation with substantial domain shifts. Next, we tested Deconomix's ability to generalize across application domains by training models on DISCO single-cell data from healthy tissue and validating them on DISCO breast cancer data. In this context, we treated cancer epithelial cells as hidden contributions and used the major cell type annotation denoted in Table S1 to reduce model complexity. An overview of the training and test data compositions is provided in Table S4. As previously, we used the Deconomix functions to generate the training and test bulks, and to evaluate the naive model, the Deconomix model, and the Deconomix+h model. The results, shown in Table S5, indicate that Deconomix models with optimized gene weights significantly outperform the naive model, achieving average correlation values of 0.68 compared to 0.37. The Deconomix+h performed similar and additionally estimated the proportions of the background contribution with $\rho = 0.564$. In summary, despite an underlying domain shift, both Deconomix models performed well, with almost all cell types exceeding $\rho = 0.6$. Notably, performance was only compromised significantly for healthy epithelial cells with performance ranging from 0.244 (naive model) and 0.262 (Deconomix), to 0.321 (Deconomix+h). This is most likely a result of the present domain shift and the binary classification of epithelial cells into healthy and cancerous, while these lie, in reality, on a gradient between healthy and cancerous.

Scenario 3: Deconomix analysis of breast cancer bulks from TCGA. Finally, we applied Deconomix to real bulk expression data from the TCGA-BRCA cohort [23, 24]. These data comprise, in total, $n = 1,176$ samples from four breast cancer subtypes: Luminal A ($n = 637$), Luminal B ($n = 233$), HER2+ ($n = 91$), and Basal-like ($n = 215$). In line with scenario 2, we established the Deconomix models on the healthy single-cell data. This has the advantage that the estimated cell-type specific gene regulation can be interpreted relative to healthy controls. Furthermore, we mitigated potential platform effects by gene-wise platform adaptation factors, and selected $p = 5,000$ genes present in both the single-cell and bulk data (see Methods). Given the comparatively complex scenario comprising potential background contributions and likely cell-type specific gene regulation between healthy single-cell and the cancerous breast cancer bulks, we applied Deconomix in model "Deconomix+h,r" (estimating hidden background and cell-type specific gene regulation). This model requires calibration of a hyperparameter λ_2 to control overfitting. Deconomix automatically screens a prespecified series of λ_2 values and finally selects the value with the lowest cross-validation error or the one according to the one standard error rule (see Methods). For the analyses presented here, we rely on the latter choice, which generally yields the more conservative model. At this point, the user already receives all results of interest: cell composition estimates, background contributions, and cell-type specific gene regulation. There are several possible downstream analysis options. First, cell compositions can be correlated with phenotypes using test strategies. Table 2 summarizes significant cell-composition differences between luminal A, luminal B, HER2+, and basal-like breast cancers obtained via a Kruskal-Wallis test, containing both raw p-values and those corrected for multiple testing according to Benjamini and Hochberg [25]. Second, cell-type specific gene regulation can be explored. We visualized the top 20 upregulated and downregulated genes across four cancer subtypes in heatmaps (Figures 3 and S2). Both coherent and subtype specific gene expression patterns were observed, as illustrated for the top 40 upregulated genes per subtype in Venn diagram Figure 4. For instance, Luminal A, Luminal B, and Her2+ share several upregulated genes, while the Basal-like subtype is more unique, with 25 genes not overlapping. Notably, five genes – B2M, IFITM1, TXN, IGHG1, and HMOX1 – are consistently upregulated across all subtypes. B2M, linked to immune regulation, is predicted to be underexpressed in epithelial cells, possibly contributing to immune escape [26]. IGHG1 is upregulated in B-cells across subtypes, in line with adaptive immune activation [27]. Third, Deconomix includes functions to perform enrichment analyses with respect to the estimated cell-type specific gene expression estimates, where both Fisher's exact test and the Mann-Whitney U-test can be used. This analysis can be performed for the individual breast cancer subtypes, and can be tailored to specific cell types of interest. For illustration, we conducted a GSEA using GSEApy [28], focusing on the MSigDB C7 database of immunologic gene sets [29, 30]. The top 15 hits, presented in Table S6, show significant enrichment in T-cell activation, regulatory T-cell signals, and innate immunity signatures, confirming the anticipated immune response in diseased tissue.

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 only cellular proportions, but also potential background contributions and cell-type specific gene regulation [7, 10, 19, 31, 32, 33]. The implementation of such complex models requires dedicated algorithmic solutions that typically also require more computing resources. Moreover, advanced methods are usually more difficult to apply and require substantial programming skills. In recent years, we established several cell-type deconvolution approaches, where we systematically addressed different analysis bottlenecks [10, 34, 35, 20], such as po-

tential background contributions, gene-regulatory adaptation, and the reliability cell composition estimates. Notably, all these bottlenecks are inevitably connected. Therefore, we combined them in Deconomix – a user-friendly Python pipeline and stand-alone GUI – for users with no or little programming experience. With Deconomix, users can rely on a highly efficient state-or-the-art pipeline for cell composition analyses that accounts for dominant analysis pitfalls. Moreover, Deconomix has rich functionality, such as functions for generating pseudo-bulk data from single-cell datasets, to perform, visualize and benchmark cell-type deconvolution, conducting hyperparameter searches for advanced models, and to conduct gene-set enrichment analysis to explore cell-type specific gene regulation.

Comprehensive software suites for cell-composition analysis are scarce and alternatives are highly needed. For instance, Omnideconv stands out as a unified framework for utilizing and benchmarking single-cell-informed cell-type deconvolution methods [36]. It currently supports 12 different approaches. In the same spirit, Decon-Benchmark offers utilities for comprehensive benchmark analysis [37]. However, in contrast to Deconomix, both lack a coherent workflow to systematically guide users through outlined challenges and analytical steps in cell-type deconvolution analysis.

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. It is further designed for users with little programming knowledge and can also be applied via a user-friendly GUI.

Methods

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 $\boldsymbol{x}_k = (x_{1k}, \ldots, x_{pk})^T \in \mathbb{R}_+^{p\times 1}$, where $k = 1, \ldots, q$ 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 $\mathbf{A}$.

Gene Weight Optimization

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 to one and assuming that there are no missing reference profiles in $\mathbf{X}$, one obtains approximately $\sum_k c_{ik} = 1$ for $i = 1, \ldots, n$ [18, 10]. For computational efficiency, we relied on non-negative least squares,

$$\hat{\mathbf{C}} = \underset{\mathbf{C}\geq 0}{\arg\min}\, \mathcal{L}(\mathbf{C}) = \underset{\mathbf{C}\geq 0}{\arg\min}\, \frac{1}{n}\, \|\mathbf{Y} - \mathbf{X}\mathbf{C}\|_F^2\,, \tag{2}$$

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 do not improve cell-type deconvolution are down-weighted. This is achieved by optimizing an outer loss function, for which we provide two different choices. In line with the Digital Tissue Deconvolution algorithm (DTD) [20], 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 [20].

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 (version 2.3.1) was used [38]. Respective functions for preparing training data and the gene-weight optimization itself are provided by Deconomix.

Inferring background contributions and cell-type-specific gene regulation factors

The algorithm Adaptive Digital Tissue Deconvolution (ADTD) [10] expands the introduced supervised loss-function learning routine with an unsupervised approach to refine deconvolution results. It utilizes the gene weights Γ determined as outlined above, 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

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

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).

Hyperparameter tuning

The Deconomix model with a hidden background (Deconomix+h) 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 $\tilde{\mathbf{c}}$ [10].

A more complex Deconomix model with a hidden background and cell-type-specific gene regulation (Deconomix+h,r) is intended to infer population estimates of cell-type-specific gene regulation alongside with the outputs described for the other models. These estimates are encoded in the adaptation factors Δ. In order to provide a computationally cheap strategy to infer Δ without extensive hyperparameter screening, we used the following strategy:

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

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

This choice ensures that genes contribute comparably to the optimization problem and that a single λ_2 value can be used for different classes.
iii. Split the data into batches. When Deconomix is used with standard settings, the data are split in five batches of approximately equal size.
iv. 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).

The former procedure provides an estimate of the generalization error together with error intervals. Deconomix provides two 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 minimum value that λ_2 can take such that it is still outside the one standard error interval obtained for λ_2 = lambda.min, in line with [39]. The latter choice was used for the presented use case analysis.

Breast cancer use case and data preprocessing

We retrieved single-cell gene expression data from healthy and cancerous breast tissue from the DISCO database [21, 22]. The obtained samples contained 34 fine-grained cell-type labels, which we aggregated into broader cell-type classes informed by known cell lineages. We mapped the original labels to a set of major cell types, featuring nine different classes, and a set of minor cell types featuring 19 classes, shown in Supplementary Table S1. We computed average expression profiles for each originally labeled cell type, which we concatenated into a reference expression matrix, comprising values for all genes and each original cell-type label. Genes were then ranked decreasingly by variance between cell types. TCGA bulk expression data [23, 24] featuring breast cancer bulk expression samples were imported and restricted to genes available in the variance-ranked list. Then, the bulk data were normalized to 10,000 counts per samples, and used to compute gene-wise mean expression values μ_{TCGA}. Simultaneously, DISCO cancer single-cell samples were selected by restricting to breast cancer samples derived from solid tumors only and labeling cells according to the minor cell type annotations denoted in Table S1. Genes were selected according to the same variance-based ordering and the data were normalized to 10,000 counts per sample. Artificial bulk mixtures were simulated from these samples, renormalized to 10,000 counts per mixture, and used to compute per-gene mean expression values μ_{DISCO}.

To account for platform-specific differences between TCGA and DISCO, we estimated gene-wise conversion factors α:

$$\alpha \approx \frac{\mu_{\text{DISCO}}}{\mu_{\text{TCGA}}}. \tag{7}$$

In absence of platform-specific effects, we expect that each gene has a similar mean expression across the two datasets. Strong deviations from $\alpha \approx 1$ primarily reflect technical rather than biological differences. Genes for which this similarity assumption was violated, were conservatively filtered by requiring $0.1 \leq \alpha \leq 10.0$, corresponding to a maximum plausible tenfold change between platforms. From the remaining genes, still ordered by variance across DISCO cell types, we retained the top 5,000 to define the final gene set.

Single-cell samples labeled as cancer epithelial cells under the major annotation, despite originating from healthy samples, were removed from the healthy subset of the DISCO data. After selecting the consensus genes and normalizing to 10,000 counts per sample, we obtained a DISCO cancer and a DISCO healthy dataset.

TCGA bulk samples were processed using the same gene set, adjusted by the estimated conversion factors, and renormalized to 10,000 counts per sample to obtain the final TCGA dataset, which represent the bulk samples to be analyzed in this case study.

Availability of source code and requirements

- Project name: Deconomix
- Project home page: `https://gitlab.gwdg.de/MedBioinf/MedicalDataScience/Deconomix`
- Package source code: `https://gitlab.gwdg.de/MedBioinf/MedicalDataScience/Deconomix/Deconomix`
- Package documentation: `https://medbioinf.pages.gwdg.de/MedicalDataScience/Deconomix/Deconomix/`
- Breast cancer use case study: `https://gitlab.gwdg.de/MedBioinf/MedicalDataScience/Deconomix/Deconomix-case-study`
- Graphical user interface: `https://gitlab.gwdg.de/MedBioinf/MedicalDataScience/Deconomix/Deconomix-gui`

- Programming language: Python >=3.11
- Licence: GPLv3

Data Availability

All datasets used in the breast cancer case study are publicly available: For the DISCO single-cell human breast atlas we retrieved the file `disco_breast_v01.h5ad` from the corresponding data repository [22]. Breast cancer bulk expression data from the TCGA-BRCA project [23] were retrieved from the GDC portal [24] via API calls available as script in the code repository of our breast cancer use case study [3]. For the gene set enrichment analysis we retrieved the ImmuneSigDB subset of C7 [29, 30] from the GSEA website as gmt-file `c7.immunesigdb.v2025.1.Hs.symbols.gmt` from following page: `https://www.gsea-msigdb.org/gsea/msigdb/human/collections.jsp`

Declarations

Competing interests

None declared.

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, 01ZX1912C, and 01EQ2407A), by the DFG SFB-TRR 274 and AL 2355/7-1 [project number: 567400630], and by Helse Vest (NCT02872259). LE and HUZ were additionally supported by zukunft.niedersachsen, the joint science funding program of the Lower Saxony Ministry of Science and Culture and the Volkswagen Foundation (project “MoReHealth”).

Authors’ contributions

Malte Mensching-Buhr: Software (lead), Methodology (supporting), Validation (supporting), Writing - review and editing (equal). **Thomas Sterr:** Methodology (lead), Software (supporting), Validation (supporting), Writing - review and editing (equal). **Dennis Völkl:** Validation (supporting), Software (supporting), Writing - review and editing (equal). **Nicole Seifert:** Writing - review and editing (equal). **Jana Tauschke:** Validation (supporting). Writing - review and editing (equal). **Laurenz Engel:** Writing - review and editing (equal). **Austin Rayford:** Validation (supporting), Writing - review and editing (equal). **Oddbjørn Straume:** Resources (equal), Writing - review and editing (equal). **Sushma N. Grellscheid:** Resources (equal), Writing - review and editing (equal). **Tim Beissbarth:** Resources (equal), Writing - review and editing (equal). **Helena U. Zacharias:** Writing - review and editing (equal). **Franziska Görtler:** Conceptualization (equal), Supervision (equal), Validation (lead), Writing - review and editing (equal). **Michael Altenbuchinger:** Conceptualization (equal), Supervision (equal), Project Administration (lead), Writing - review and editing (equal).

Acknowledgements

Malte Mensching-Buhr was supported by the Ph.D. program "Genome Science" - International Max Planck Research School.

Model	Algorithm	Gene Weights	Inferred outputs			
			C	**c**	**x**	Δ
Naive	NMF with constant feature matrix	–	✓	–	–	–
Deconomix	DTD	✓	✓	–	–	–
Deconomix+h	DTD + ADTD with $\lambda_1 \to \infty$, $\lambda_2 \to \infty$	✓	✓	✓	✓	–
Deconomix+h,r	DTD + ADTD with $\lambda_1 \to \infty$, λ_2 via hyperparameter search	✓	✓	✓	✓	✓

Table 1. Overview of the Deconomix models used in the breast cancer use case study, summarizing the underlying algorithms, the use of gene weight optimization, and the quantities inferred by each model: Cellular contributions of known cell types **C**, cellular contributions of a hidden background **c**, consensus expression profile of the hidden background **x** and cell-type-specific gene regulation factors Δ.

Cell type	Luminal A	Luminal B	HER2+	Basal-like	H-statistic	p-value	p-value-corrected
B cell	0.059 ± 0.086	0.054 ± 0.095	0.116 ± 0.167	0.102 ± 0.165	16.257	0.001	0.0011
CD4 T cell	0.129 ± 0.087	0.080 ± 0.081	0.059 ± 0.075	0.046 ± 0.072	202.173	1.43e-43	7.15e-43
CD8 T cell	0.059 ± 0.070	0.111 ± 0.081	0.050 ± 0.063	0.048 ± 0.073	117.079	3.28e-25	1.09e-24
Endothelial cell	0.073 ± 0.039	0.083 ± 0.045	0.063 ± 0.048	0.048 ± 0.041	94.752	2.09e-20	3.48e-20
Fibroblast	0.078 ± 0.053	0.067 ± 0.051	0.053 ± 0.058	0.062 ± 0.088	61.675	2.58e-13	3.68e-13
Healthy Epithelial	0.270 ± 0.112	0.242 ± 0.098	0.321 ± 0.117	0.362 ± 0.168	108.894	1.9e-23	4.74e-23
Myeloid cell	0.094 ± 0.064	0.081 ± 0.061	0.096 ± 0.058	0.112 ± 0.085	16.064	0.0011	0.0011
NK cell	0.071 ± 0.064	0.087 ± 0.065	0.110 ± 0.093	0.142 ± 0.096	97.060	6.66e-21	1.33e-20
Perivascular cell	0.021 ± 0.027	0.022 ± 0.024	0.011 ± 0.020	0.016 ± 0.027	41.371	5.46e-09	6.82e-09
hidden	0.146 ± 0.087	0.172 ± 0.078	0.122 ± 0.185	0.061 ± 0.121	329.247	4.65e-71	4.65e-70

Table 2. Average cellular composition of four breast cancer subtypes within the TCGA-BRCA cohort [23, 24] estimated by a Deconomix+h,r model which infers hidden background contributions and cell-type specific gene regulation. Values denoted as mean $\pm$ one standard deviation. H-statistic and p-values determined by Kruskal-Wallis tests to identify significant differences in the composition of the different breast cancer subtypes. P-values adjusted for multiple testing by controlling the false-discovery rate (Benjamini-Hochberg [25]).

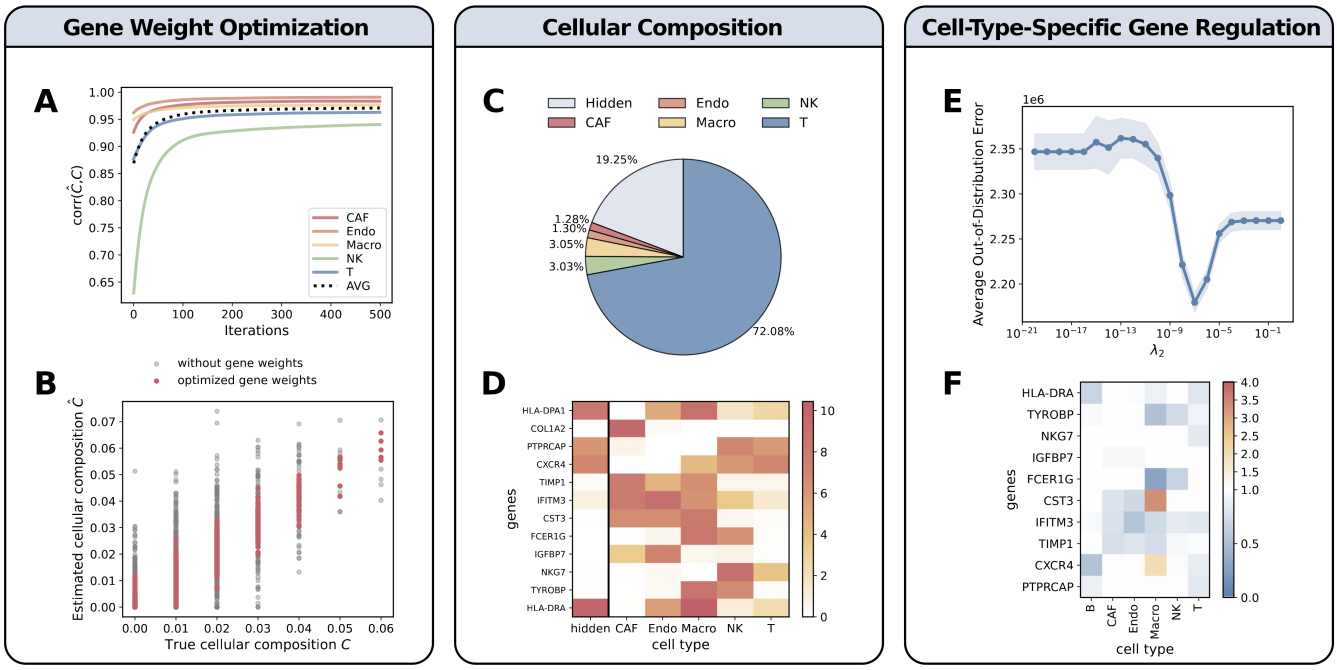

Figure 1. Deconomix Analysis Modules with example realizations: (A) Average and cell-type specific correlation between inferred cellular proportions and respective ground-truth values (y-axis) versus the iteration of the Deconomix optimization procedure (x-axis). (B) Scatter plot contrasting predictions (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 profile (column "hidden"). (E) Hyper-parameter grid search to infer cell-type specific gene regulation and (F) the corresponding cell-type specific gene regulation matrix.

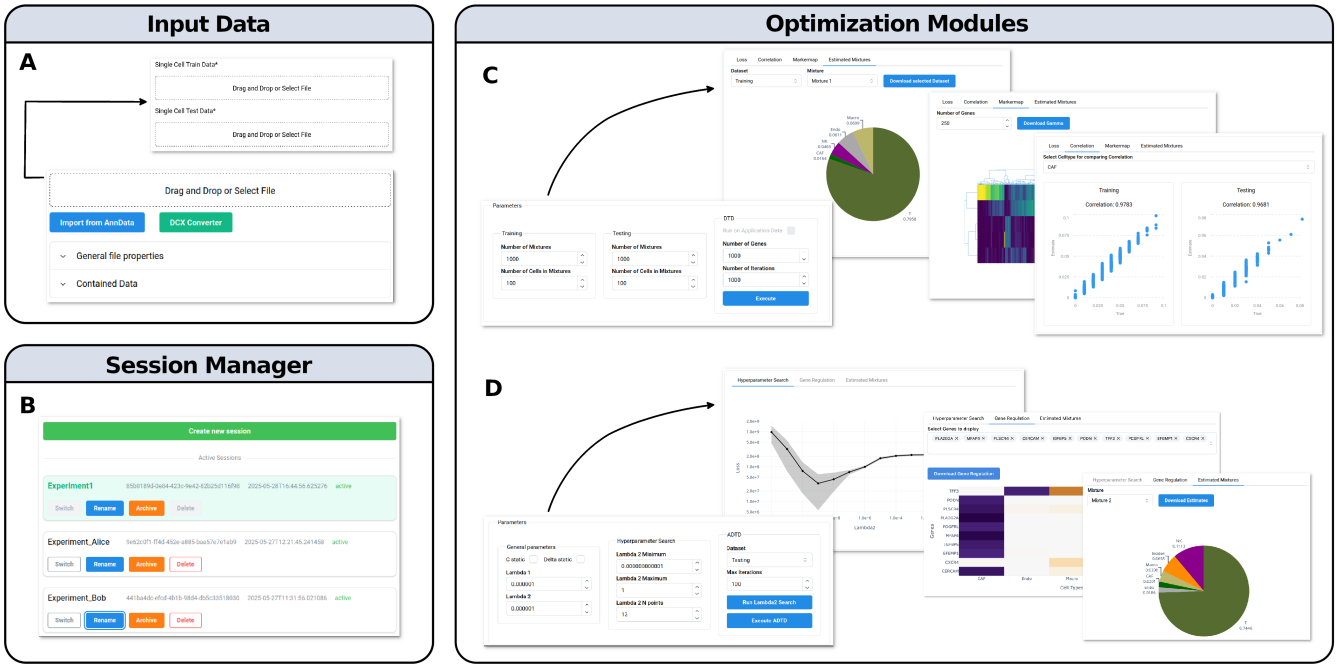

Figure 2. Deconomix Graphical User Interface: The application is locally hosted and enables the users to work with sensitive data. Through the session manager (B) multiple users can work within the same interface. After integrating the data through the interface via .h5ad or .dcx (internal file format; conversion tool can be found on the website) (A) the user can run different Deconomix models (C,D) and the exportable results are visualized with interactive plots, including correlation plots to ground truth values, pie charts of estimated cellular compositions, heatmaps for reference and learned expression profiles, and heatmaps for cell-type-specific gene regulation factors.

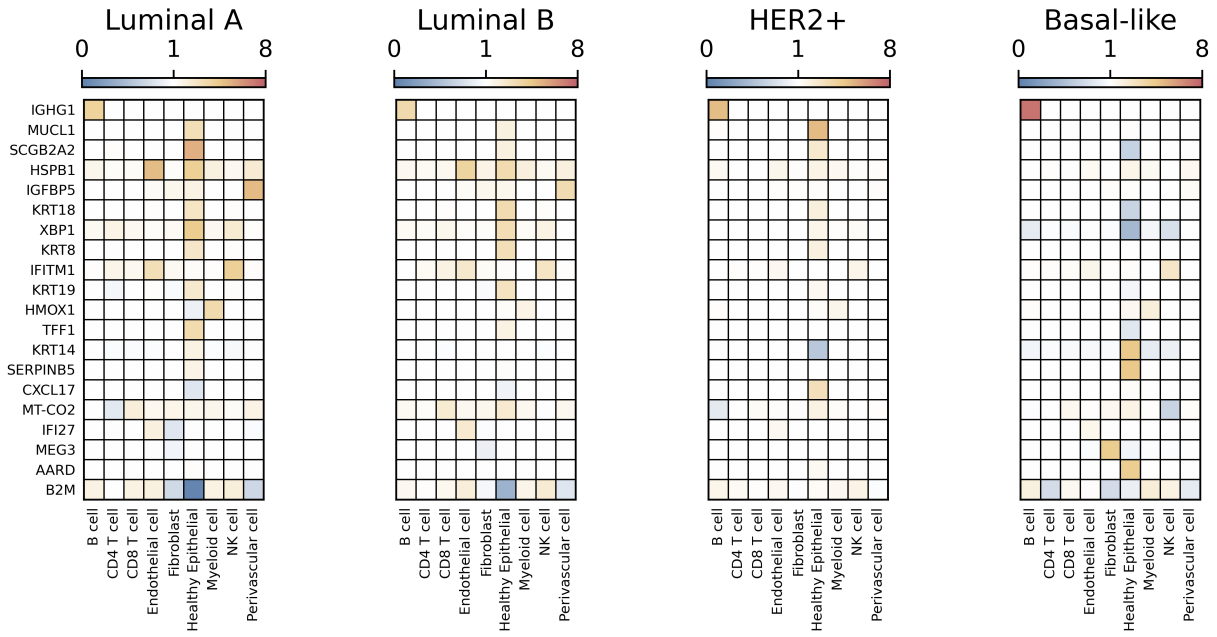

Figure 3. Heatmap of the 20 genes most upregulated across the four breast cancer subtypes. Values between 0 and 1 (blue) correspond to down-regulation, white to no regulation, values above 1 (yellow over orange and red) correspond to positive gene regulation.

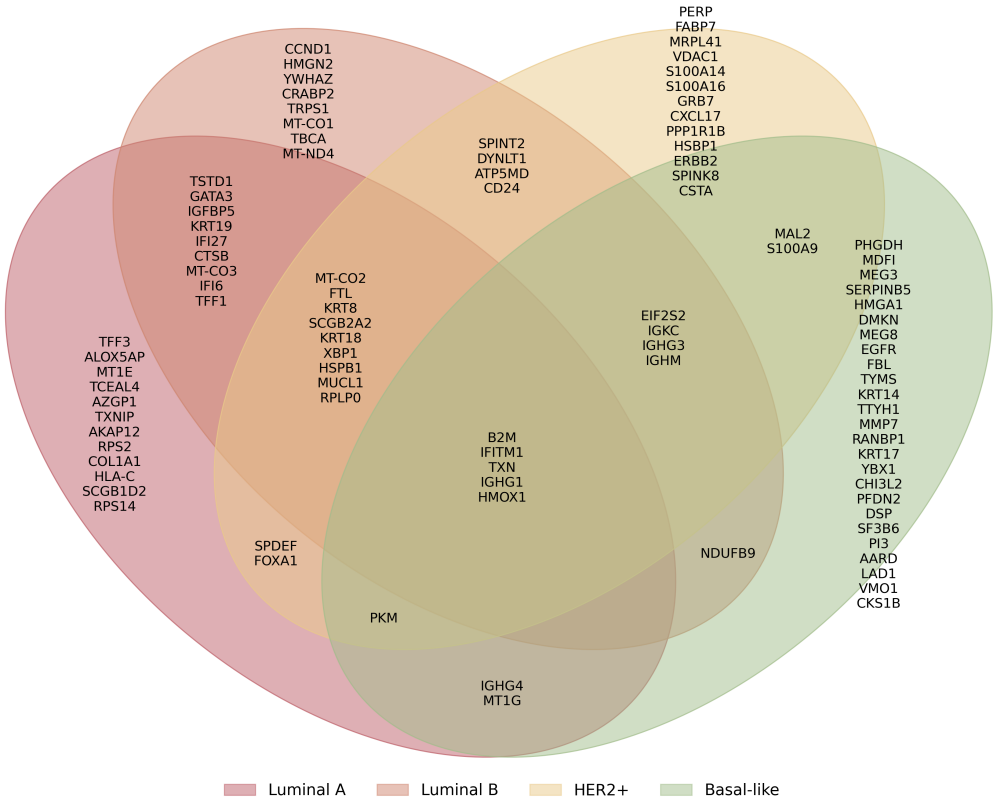

Figure 4. Venn diagram illustrating the overlap of the 40 top up-regulated genes in any cell type across the four breast cancer subtypes within TCGA breast cancer data [23, 24], inferred by a Deconomix+h,r model with λ_2 = 1e-11.

Supplementary Material

Cell type (DISCO)	Cell type (minor)	Cell type (major)
Arterial EC	Arterial EC	Endothelial cell
B cell	B cell	B cell
Breast basal cell	Healthy Epithelial	Healthy Epithelial
Breast cancer specific luminal cell	Cancer Epithelial	Cancer Epithelial
Breast cancer specific proliferation luminal cell	Cancer Epithelial	Cancer Epithelial
CD4 T	CD4 T cell	CD4 T cell
CFD fibroblast	Fibroblast	Fibroblast
CXCL1/2/3 fibroblast	Fibroblast	Fibroblast
CXCL13 exhausted CD8 T	CD8 T cell	CD8 T cell
Capillary EC	Capillary EC	Endothelial cell
GZMH CD8 T	CD8 T cell	CD8 T cell
GZMK CD8 T	CD8 T cell	CD8 T cell
INF responsed T	CD4 T cell	CD4 T cell
IgA plasma	IgA plasma	B cell
IgG plasma cell	IgG plasma	B cell
Luminal progenitor	Healthy Epithelial	Healthy Epithelial
Lymphatic EC	Lymphatic EC	Endothelial cell
Macrophage	Macrophage	Myeloid cell
Mast cell	Granulocyte	Myeloid cell
Monocyte	Monocyte	Myeloid cell
NK cell	NK cell	NK cell
POSTN fibroblast	Fibroblast	Fibroblast
Pericyte	Pericyte	Perivascular cell
Proliferation T/NK	NK cell	NK cell
Proliferation macrophage	Macrophage	Myeloid cell
Smooth muscle cell	Smooth muscle cell	Perivascular cell
TGM2 luminal cell	Healthy Epithelial	Healthy Epithelial
TNBC-specific epithelial cell	Cancer Epithelial	Cancer Epithelial
Tfh	CD4 T cell	CD4 T cell
Treg	CD4 T cell	CD4 T cell
Venous EC	Venous EC	Endothelial cell
cDC2	Dendritic cell	Myeloid cell
mregDC	Dendritic cell	Myeloid cell
pDC	Dendritic cell	Myeloid cell

Table S1. Aggregation of the original, fine-grained cell-type labels from the DISCO single-cell dataset [22] into two sets of more coarse labels. The minor subset features 19, and the major subset nine distinct classes.

Cell type	Train number	Train percentage	Test number	Test percentage
Arterial EC	335	0.555	346	0.351
B cell	2524	4.184	2505	2.538
CD4 T cell	13376	22.175	13382	13.559
CD8 T cell	10012	16.598	9971	10.103
Capillary EC	977	1.620	1051	1.065
Dendritic cell	1679	2.783	1571	1.592
Fibroblast	8023	13.301	7990	8.095
Granulocyte	449	0.744	439	0.445
Healthy Epithelial	5049	8.370	5171	5.239
IgA plasma	568	0.942	570	0.578
IgG plasma	2285	3.788	2197	2.226
Lymphatic EC	155	0.257	169	0.171
Macrophage	7968	13.210	7901	8.005
Monocyte	1112	1.844	1128	1.143
NK cell	2559	4.242	2736	2.772
Pericyte	475	0.787	469	0.475
Smooth muscle cell	1497	2.482	1477	1.496
Venous EC	1277	2.117	1276	1.293
sum immune cell types	60320	100.000	60349	61.146
hidden contribution	-	-	38348	38.854
total	60320	100.000	98697	100.000

Table S2. Composition of single-cell samples in a random test/train split of the cancer samples within the DISCO dataset [22]. Relabeled using the minor cell-type labels defined in Table S1. Cancer epithelial cells were dropped from the training set but preserved as 'hidden contribution' in the test set.

Cell Type	Train Naive	Train Deconomix	Test Naive	Test Deconomix	Test Deconomix+h
Arterial EC	0.202	0.672	0.047 ± 0.031	0.427 ± 0.036	0.523 ± 0.03
B cell	0.282	0.848	0.185 ± 0.04	0.793 ± 0.014	0.8 ± 0.016
CD4 T cell	0.470	0.843	0.414 ± 0.014	0.813 ± 0.013	0.768 ± 0.015
CD8 T cell	0.409	0.756	0.319 ± 0.022	0.685 ± 0.023	0.685 ± 0.023
Capillary EC	0.257	0.708	0.171 ± 0.027	0.595 ± 0.019	0.636 ± 0.019
Dendritic cell	0.318	0.745	0.038 ± 0.031	0.55 ± 0.02	0.622 ± 0.012
Fibroblast	0.668	0.942	0.663 ± 0.017	0.912 ± 0.004	0.914 ± 0.004
Granulocyte	0.541	0.743	0.463 ± 0.041	0.59 ± 0.031	0.627 ± 0.032
Healthy Epithelial	0.725	0.938	0.356 ± 0.031	0.345 ± 0.028	0.473 ± 0.027
IgA plasma	0.777	0.879	0.788 ± 0.014	0.835 ± 0.012	0.83 ± 0.012
IgG plasma	0.852	0.945	0.852 ± 0.012	0.928 ± 0.005	0.893 ± 0.01
Lymphatic EC	0.246	0.689	0.207 ± 0.037	0.459 ± 0.029	0.481 ± 0.026
Macrophage	0.564	0.809	0.495 ± 0.03	0.756 ± 0.023	0.784 ± 0.02
Monocyte	0.166	0.412	0.104 ± 0.016	0.248 ± 0.037	0.289 ± 0.034
NK cell	0.224	0.629	0.137 ± 0.02	0.392 ± 0.024	0.401 ± 0.025
Pericyte	0.202	0.607	0.148 ± 0.04	0.522 ± 0.03	0.519 ± 0.026
Smooth muscle cell	0.490	0.846	0.412 ± 0.025	0.801 ± 0.014	0.814 ± 0.014
Venous EC	0.318	0.812	0.223 ± 0.032	0.768 ± 0.013	0.791 ± 0.011
mean (excl. hidden)	0.428	0.768	0.335 ± 0.027	0.634 ± 0.021	0.658 ± 0.02
hidden	-	-	-	-	0.655 ± 0.025
hidden reference	-	-	-	-	0.667 ± 0.001

Table S3. Performance assessment of cell-type deconvolution on bulk profiles simulated from the breast cancer samples within the DISCO dataset[22]. Calculated as Spearman's correlation coefficient ρ between true and estimated cellular compositions. Performance on training set is shown to illustrate training success. Evaluation on a different test set was conducted on ten folds of the data and is denoted as mean $\pm$ one standard deviation. Three different models are featured: A naive model without Deconomix' gene weight optimization, a Deconomix model, and a Deconomix+h model, which performs hidden background estimation.

Cell type	Train number	Train percentage	Test number	Test percentage
B cell	269	0.986	10649	5.408
CD4 T cell	3642	13.348	26758	13.588
CD8 T cell	4792	17.563	19983	10.148
Endothelial cell	5504	20.173	5586	2.837
Fibroblast	5311	19.466	16013	8.132
Healthy Epithelial	1979	7.253	10220	5.190
Myeloid cell	3564	13.063	22247	11.298
NK cell	993	3.639	5295	2.689
Perivascular cell	1230	4.508	3918	1.990
sum immune cell types	27284	100.000	120669	61.280
hidden contribution	-	-	76249	38.721
total	27284	100.000	196918	100.000

Table S4. Composition of single-cell samples in a test/train split of healthy samples versus cancer samples within the DISCO dataset [22]. Relabeled using the major cell-type labels defined in Table S1.

Cell Type	Train Naive	Train Deconomix	Test Naive	Test Deconomix	Test Deconomix+h
B cell	0.581	0.826	0.599 ± 0.026	0.87 ± 0.009	0.836 ± 0.01
CD4 T cell	0.245	0.808	-	0.758 ± 0.013	0.752 ± 0.011
CD8 T cell	0.241	0.798	0.005 ± 0.024	0.617 ± 0.029	0.601 ± 0.029
Endothelial cell	0.753	0.960	0.341 ± 0.019	0.792 ± 0.009	0.794 ± 0.011
Fibroblast	0.785	0.970	0.670 ± 0.022	0.857 ± 0.008	0.861 ± 0.01
Healthy Epithelial	0.733	0.942	0.244 ± 0.025	0.262 ± 0.025	0.321 ± 0.03
Myeloid cell	0.884	0.957	0.744 ± 0.015	0.924 ± 0.003	0.919 ± 0.003
NK cell	0.271	0.750	0.183 ± 0.036	0.398 ± 0.02	0.359 ± 0.024
Perivascular cell	0.680	0.912	0.196 ± 0.028	0.662 ± 0.02	0.660 ± 0.019
mean (excl. hidden)	0.575	0.880	0.373 ± 0.024	0.682 ± 0.015	0.678 ± 0.016
hidden	-	-	-	-	0.564 ± 0.016
hidden reference	-	-	-	-	0.550 ± 0.001

Table S5. Performance assessment of different cell-type deconvolution models established on healthy single-cell data and applied bulk profiles simulated from breast cancer samples both within the DISCO dataset [22]. Calculated as Spearman's correlation coefficient ρ between true and estimated cellular compositions. Performance on training set is shown to illustrate training success. Evaluation on a different test set was conducted on ten folds of the data and is denoted as mean $\pm$ one standard deviation. Three different models are featured: A naive model without Deconomix' gene weight optimization, a Deconomix model, and a Deconomix+h model, which performs hidden background estimation. The naive model estimated the abundance of CD4 T cells in the test set to be zero which is why no correlation coefficient could be calculated for that field.

Term	NES	FDR q-val
GSE24574_BCL6_LOW_TFH_VS_TCONV_CD4_TCELL_DN	3.1449646	0.0000857
GSE15930_NAIVE_VS_48H_IN_VITRO_STIM_IL12_CD8_TCELL_DN	3.1274405	0.0000857
GSE13547_2H_VS_12_H_ANTI_IGM_STIM_BCELL_DN	3.0852316	0.0000857
GSE14415_INDUCED_TREG_VS_FAILED_INDUCED_TREG_UP	3.1502211	0.0001714
GSE36476_CTRL_VS_TSST_ACT_72H_MEMORY_CD4_TCELL_YOUNG_DN	2.9092610	0.0004163
GSE15930_NAIVE_VS_48H_IN_VITRO_STIM_CD8_TCELL_DN	2.9321779	0.0004285
GSE19888_ADENOSINE_A3R_ACT_VS_TCELL_MEMBRANES_ACT_AND_A3R_INH_PRETREAT_IN_MAST_CELL_DN	2.9187470	0.0004571
GSE13547_CTRL_VS_ANTI_IGM_STIM_BCELL_12H_UP	2.8839113	0.0005035
GSE2405_0H_VS_9H_A_PHAGOCYTOPHILUM_STIM_NEUTROPHIL_DN	2.7845614	0.0013322
GSE28726_NAIVE_CD4_TCELL_VS_NAIVE_VA24NEG_NKTCELL_UP	2.7777230	0.0013426
GSE2405_0H_VS_24H_A_PHAGOCYTOPHILUM_STIM_NEUTROPHIL_UP	2.7855981	0.0014397
GSE36476_CTRL_VS_TSST_ACT_40H_MEMORY_CD4_TCELL_YOUNG_DN	2.7933093	0.0014759
GSE14415_INDUCED_TREG_VS_TCONV_UP	2.7577552	0.0014833
GSE25088_WT_VS_STAT6_KO_MACROPHAGE_IL4_STIM_DN	2.7157811	0.0022527
GSE24634_TREG_VS_TCONV_POST_DAY5_IL4_CONVERSION_UP	2.7091128	0.0023082

Table S6. Gene set enrichment analysis against MSigDB C7 (immunlogic signature gene sets) [29, 30] for the inferred gene regulation factors across all cancer subtypes and cell types from the breast cancer use case study: Top 15 gene sets with significant enrichment (FDR q-value < 0.05) ranked by normalized enrichment score (NES).

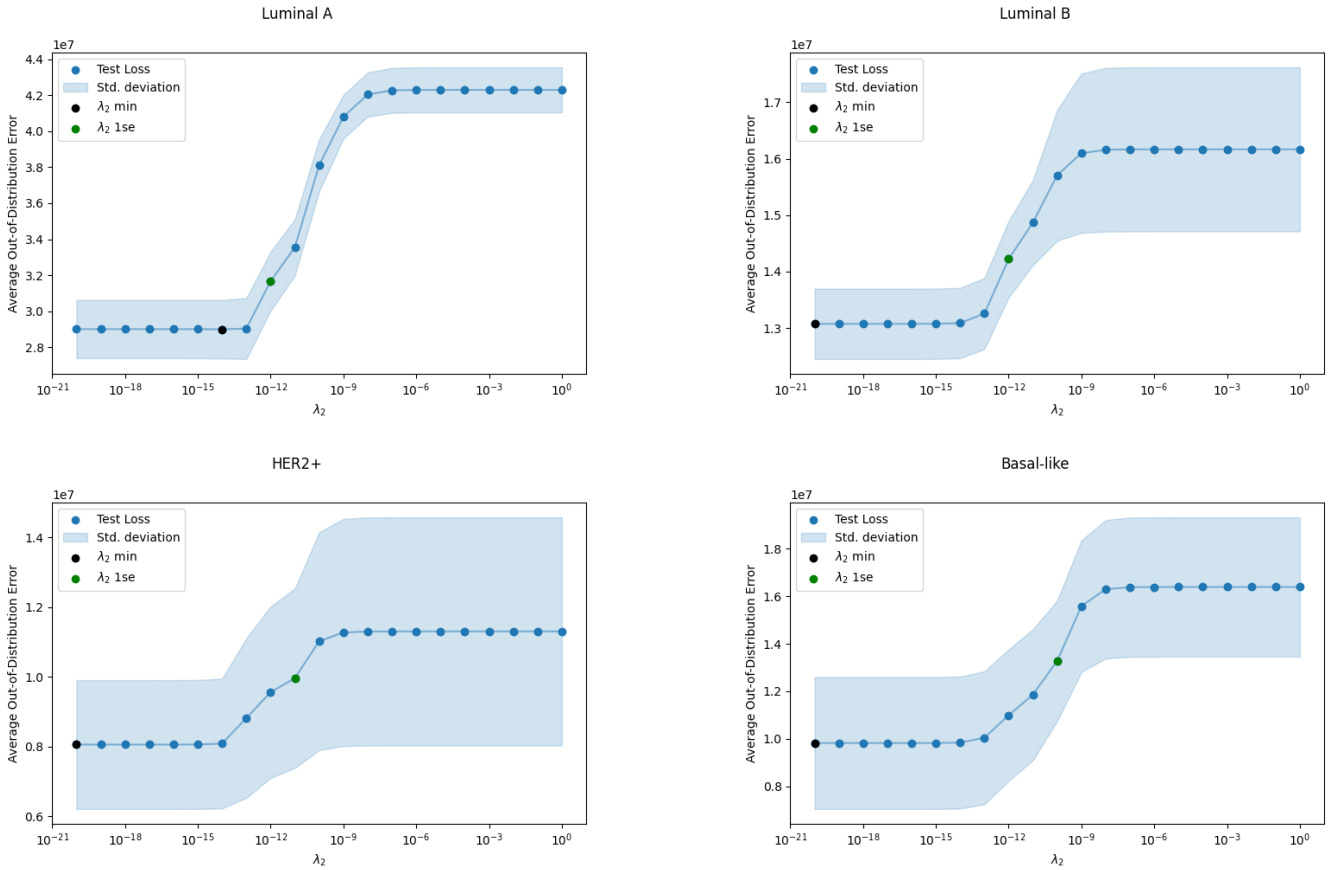

Figure S1. Hyperparameter search for the Deconomix+h,r model on four subsets of different breast cancer subtypes within the TCGA bulk dataset [23, 24]. Blue dots represent mean values of error on a left-out test set during five-fold cross-validation. Light blue regions represent one standard deviation around those mean values. Black dots represent the minimal λ_2, while the green dots denote the optimal λ_2 recommended by the one-standard-error rule (1SE rule).

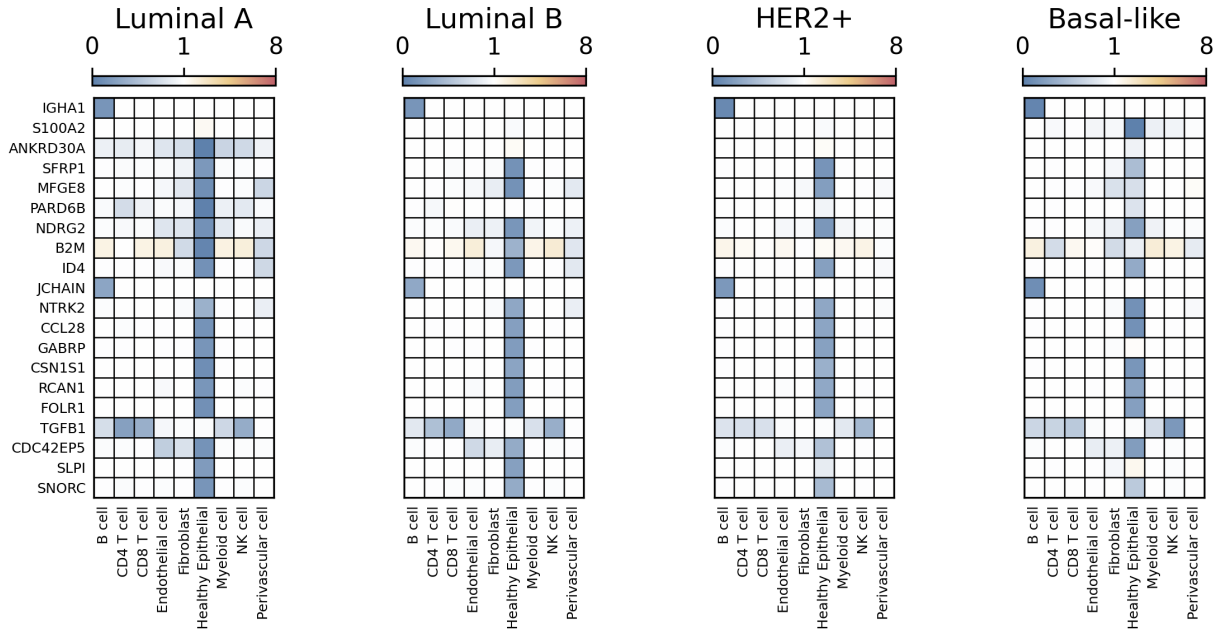

Figure S2. Heatmap of the 20 genes most downregulated across the four breast cancer subtypes. Values between 0 and 1 (blue) correspond to down-regulation, white to no regulation, values above 1 (yellow over orange and red) correspond to positive gene regulation.

References

1. Mensching-Buhr M, Sterr T, Völkl D, Görtler F, Altenbuchinger M, Deconomix Python Package;. https://gitlab.gwdg.de/MedBioinf/MedicalDataScience/Deconomix/Deconomix/-/releases/1.9.0.
2. Völkl D, Mensching-Buhr M, Sterr T, Deconomix Graphical User Interface;. https://gitlab.gwdg.de/MedBioinf/MedicalDataScience/Deconomix/Deconomix-gui/-/releases/1.0.0.
3. Mensching-Buhr M, Sterr T, Deconomix Case Study;. https://gitlab.gwdg.de/MedBioinf/MedicalDataScience/Deconomix/Deconomix-case-study/-/releases/1.0.0.
4. Becker WR, Nevins SA, Chen DC, Chiu R, Horning AM, Guha TK, et al. Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. Nature Genetics 2022;54(7):985–995. https://doi.org/10.1038/s41588-022-01088-x.
5. Wang Y, Chen D, Liu Y, Shi D, Duan C, Li J, et al. Multidirectional characterization of cellular composition and spatial architecture in human multiple primary lung cancers. Cell Death & Disease 2023;14(7):462. https://doi.org/10.1038/s41419-023-05992-w.
6. Cook DP, Vanderhyden BC. Ovarian cancer and the evolution of subtype classifications using transcriptional profiling†. Biology of Reproduction 2019 06;101(3):645–658. https://doi.org/10.1093/biolre/ioz099.
7. Chu T, Wang Z, Pe'er D, Danko CG. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. Nature Cancer 2022;3(4):505–517. https://doi.org/10.1038/s43018-022-00356-3.
8. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. In: Cancer systems biology Springer; 2018.p. 243–259.
9. Menden K, Marouf M, Oller S, Dalmia A, Magruder DS, Kloiber K, et al. Deep learning–based cell composition analysis from tissue expression profiles. Science Advances 2020 Jul;6(30). https://doi.org/10.1126/sciadv.aba2619.
10. Görtler F, Mensching-Buhr M, Skaar Schrod S, Sterr T, Schäfer A, et al. Adaptive digital tissue deconvolution. Bioinformatics 2024 06;40(Supplement_1):i100–i109. https://doi.org/10.1093/bioinformatics/btae263.
11. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nature communications 2020;11(1):1–14.
12. Sutton GJ, Poppe D, Simmons RK, Walsh K, Nawaz U, Lister R, et al. Comprehensive evaluation of deconvolution methods for human brain gene expression. Nature Communications 2022;13(1):1358.
13. Chen J, Liu W, Luo T, Yu Z, Jiang M, Wen J, et al. A comprehensive comparison on cell-type composition inference for spatial transcriptomics data. Briefings in Bioinformatics 2022;23(4):bbac245.
14. Wei J, Mayberry CL, Lv X, Hu F, Khan T, Logan NA, et al. IL3-Driven T Cell–Basophil Crosstalk Enhances Antitumor Immunity. Cancer Immunology Research 2024 May;12(7):822–839. http://dx.doi.org/10.1158/2326-6066.CIR-23-0851.
15. Ivich A, Davidson NR, Grieshober L, Li W, Hicks SC, Doherty JA, et al. Missing cell types in single-cell references impact deconvolution of bulk data but are detectable. bioRxiv 2024;https://www.biorxiv.org/content/early/2024/04/28/2024.04.25.590992.
16. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference;10(1):380. https://www.nature.com/articles/s41467-018-08023-x.
17. Schelker M, Feau S, Du J, Ranu N, Klipp E, MacBeath G, et al. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nature communications 2017;8(1):1–12.
18. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. eLife 2017 Nov;6. https://doi.org/10.7554/elife.26476.
19. Simeth J, Hüttl P, Schön M, Nozari Z, Huttner M, Schmidt T, et al. Virtual Tissue Expression Analysis. bioRxiv 2023;https://www.biorxiv.org/content/early/2023/11/17/2023.11.16.567357.
20. Görtler F, Schoen M, Simeth J, Solbrig S, Wettig T, Oefner PJ, et al. Loss-Function Learning for Digital Tissue Deconvolution. Journal of Computational Biology 2020;27(3):342–355.
21. Li M, Zhang X, Ang KS, Ling J, Sethi R, Lee NYS, et al. DISCO: a database of Deeply Integrated human Single-Cell Omics data. Nucleic Acids Research 2021;50(D1):D596–D602. http://dx.doi.org/10.1093/nar/gkab1020.
22. Li M, DISCO tissue atlas h5ad version1. Zenodo;. https://zenodo.org/records/7396984.
23. Lingle W, Erickson BJ, Zuley ML, Jarosz R, Bonaccio E, Filippini J, et al., The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA). The Cancer Imaging Archive;. https://www.cancerimagingarchive.net/collection/tcga-brca/.
24. Heath AP, Ferretti V, Agrawal S, An M, Angelakos JC, Arya R, et al. The NCI Genomic Data Commons;53(3):257–262. https://www.nature.com/articles/s41588-021-00791-5.
25. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing;57(1):289–300. https://www.jstor.org/stable/2346101.
26. Wang H, Liu B, Wei J. Beta2-microglobulin(B2M) in cancer immunotherapies: Biological function, resistance and remedy;517:96–104. https://www.sciencedirect.com/science/article/pii/S0304383521002937.
27. Bonilla FA, Oettgen HC. Adaptive immunity;125(2):S33–S40. https://www.sciencedirect.com/science/article/pii/S0091674909014055.
28. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python;39(1):btac757. https://doi.org/10.1093/bioinformatics/btac757.
29. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles;102(43):15545–15550. https://www.pnas.org/doi/10.1073/pnas.0506580102.
30. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0;27(12):1739–1740. https://doi.org/10.1093/bioinformatics/btr260.
31. Chen ZJ, Rahmani E, Halperin E. Unico: a unified model for cell-type resolution genomics from heterogeneous omics data;26(1):333. https://doi.org/10.1186/s13059-025-03776-3.

32. Tang C, Sun Q, Zeng X, Li G, Yang X, Liu F, et al. Cell type-specific inference from bulk RNA-sequencing data by integrating single-cell reference profiles via EPIC-unmix;26(1):395. https://doi.org/10.1186/s13059-025-03847-5.
33. Wang J, Roeder K, Devlin B. Bayesian estimation of cell type–specific gene expression with prior derived from single-cell data;31(10):1807–1818. http://genome.cshlp.org/content/31/10/1807.
34. Völkl D, Mensching-Buhr M, Sterr T, Bolz S, Schäfer A, Seifert N, et al. HIDE: hierarchical cell-type deconvolution;41:i207–i216. https://doi.org/10.1093/bioinformatics/btaf179.
35. Seifert N, Engel L, Tauschke J, Sterr T, Demmer J, Mensching-Buhr M, et al., DynaMiCs – Dynamic cell-type deconvolution ensembles for Mapping in mixed Conditions. bioRxiv;. https://www.biorxiv.org/content/10.1101/2025.05.08.652868v2, pages: 2025.05.08.652868 Section: New Results.
36. Dietrich A, Merotto L, Pelz K, Eder B, Zackl C, Reinisch K, et al. omnideconv: a unifying framework for using and benchmarking single-cell-informed deconvolution of bulk RNA-seq data;27(1):6. https://doi.org/10.1186/s13059-026-03955-w.
37. Nguyen H, Nguyen H, Tran D, Draghici S, Nguyen T. Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges;52(9):4761–4783. https://doi.org/10.1093/nar/gkae267.
38. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems 32 Curran Associates, Inc.; 2019.p. 8024–8035.
39. Friedman J, Tibshirani R, Hastie T. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 2010;33(1):1–22.