OXFORD

Briefings in Bioinformatics, 2025, **26(3)**, bbaf234

https://doi.org/10.1093/bib/bbaf234

Problem Solving Protocol

DECEPTICON: a correlation-based strategy for RNA-seq deconvolution inspired by a variation of the Anna Karenina principle

Fulan Deng[1,2,‡], Jiawei Zou[3,‡], Miaochen Wang[4,‡], Yida Gu[5], Jiale Wu[6], Lianchong Gao[7], Yuan Ji[8], Henry H.Y. Tong[2], Jie Chen[9], Wantao Chen[10], Lianjiang Tan[1], Yaoqing Chu[1,*], Xin Zou[11,12,*], Jie Hao[13,14,*]

[1]School of Materials Science and Engineering, Shanghai Institute of Technology, 100 Haiquan Road, Fengxian District, Shanghai 201418, China
[2]Centre for Artificial Intelligence Driven Drug Discovery, Faculty of Applied Sciences, Macao Polytechnic University, Rua de Luís Gonzaga Gomes, Macao SAR 999078, China
[3]Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, University of Chinese Academy of Sciences, 320 Yueyang Road, Xuhui District, Shanghai 200031, China
[4]Department of General Dentistry, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, College of Stomatology, Shanghai Jiao Tong University, 1908 Gaoke West Road, Pudong New District, Shanghai 200240, China
[5]Guangdong Provincial/Zhuhai Key Laboratory of Interdisciplinary Research and Application for Data Science, Beijing Normal-Hong Kong Baptist University, 2000 Jintong Road, Tangjiawan, Xiangzhou District, Zhuhai 519087, China
[6]Mathematics and Science College, Shanghai Normal University, 100 Guilin Road, Xuhui District, Shanghai 200233, China
[7]Shanghai Centre for Systems Biomedicine, Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Centre for Systems Biomedicine, Shanghai Jiao Tong University, 800 Dong Chuan Road, Minhang District, Shanghai 200240, China
[8]Molecular Pathology Center, Department of Pathology, Zhongshan Hospital, Fudan University, 966 Huaihai Middle Road, Xuhui District, Shanghai 200032, China
[9]Center for Ultrafast Science and Technology, Key Laboratory for Laser Plasmas (Ministry of Education), School of Physics and Astronomy, Collaborative Innovation Center of IFSA (CICIFSA, Shanghai Jiao Tong University, 800 Dongchuan Road, Minhang District, Shanghai 200240, China
[10]Ninth People's Hospital, Shanghai Key Laboratory of Stomatology & Shanghai Research Institute of Stomatology, National Clinical Research Center of Stomatology, Shanghai Jiao Tong University School of Medicine, Huangpu District, Shanghai 200011, China
[11]School of Medicine, Linyi University, Shuangling Road, Lanshan District, Linyi, Shandong 276000, China
[12]Digital Diagnosis and Treatment Innovation Center for Cancer, Institute of Translational Medicine, Shanghai Jiao Tong University, 800 Dong Chuan Road, Minhang District, Shanghai 200240, China
[13]Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Botanical Garden, 3888 Chenhua Road, Songjiang District, Shanghai 201602, China
[14]Institute of Clinical Science, Zhongshan Hospital, Fudan University, No.180 Fenglin Road, Xuhui District, Shanghai 200032, China

*Corresponding authors. Yaoqing Chu, School of Materials Science and Engineering, Shanghai Institute of Technology, 100 Haiquan Road, Fengxian District, Shanghai 201418, China. E-mail: yq_chu@sit.edu.cn; Xin Zou, School of Medicine, Linyi University, Shuangling Road, Lanshan District, Linyi, Shandong 276000, China. E-mail: albumxin@qq.com; Jie Hao, Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Botanical Garden, 3888 Chenhua Road, Songjiang District, Shanghai 201602, China. E-mail: jhao@fudan.edu.cn

‡Fulan Deng, Jiawei Zou, and Miaochen Wang contributed equally.

Abstract

Accurately deconvoluting cellular composition from bulk RNA-seq data is pivotal for understanding the tumor microenvironment and advancing precision medicine. Existing methods often struggle to consistently and accurately quantify cell types across heterogeneous RNA-seq datasets, particularly when ground truths are unavailable. In this study, we introduce DECEPTICON, a deconvolution strategy inspired by the Anna Karenina principle, which postulates that successful outcomes share common traits, while failures are more varied. DECEPTICON selects top-performing methods by leveraging correlations between different strategies and combines them dynamically to enhance performance. Our approach demonstrates superior accuracy in predicting cell-type proportions across multiple tumor datasets, improving correlation by 23.9% and reducing root mean square error by 73.5% compared to the best of 50 analyzed strategies. Applied to The Cancer Genome Atlas (TCGA) datasets for breast carcinoma, cervical squamous cell carcinoma, and lung adenocarcinoma, DECEPTICON-based predictions showed improved differentiation between patient prognoses. This correlation-based strategy offers a reliable, flexible tool for deconvoluting complex transcriptomic data and highlights its potential in refining prognostic assessments in oncology and advancing cancer biology.

Keywords: scRNA-seq data; bulk RNA-seq; deconvolution; AKP

Introduction

Tumor tissue is a complex microenvironment consisting of various cell types, including tumor cells, stromal cells, fibroblasts, and immune cells [1, 2]. Previous studies have revealed that the tumor microenvironment is closely related to the prognosis and efficacy of cancer treatment. Tumor-infiltrating immune cells play a central role in tumor control and therapeutic response [3–5]. The quantitative analysis of immune cells can reveal

Received: November 18, 2024. **Revised:** February 22, 2025. **Accepted:** April 29, 2025

prognostic markers and help to elucidate the mechanisms of treatment response in various tumor types [6–9]. In addition, in the biomedical field, researchers use the deconvolution approach to extract information about cell composition from bulk RNA-seq data. This technique has broad applications in various biological processes, such as disease [10], development [11], aging [12], and cancer immunology [6–9, 13, 14].

Methods such as fluorescence-activated cell sorting (FACS), immunohistochemistry staining, and flow cytometry have been used as the gold standard for estimating immune cell contents [15–17]. These targeted methods are often limited by feasibility or cost and cannot be universally applied. Therefore, the potential of computational methods as an alternative approach to quantify immune cells from bulk RNA-seq samples has been recognized. Multiple methods have been developed to deconvolute bulk RNA-seq data, such as MCP-counter [18], xCell [19], and CIBERSORT [20]. [21–27]. Several studies have benchmarked various deconvolution methods for quantifying cell types from different perspectives. Sturm *et al.* [28] were the first to systematically and quantitatively compare different methods for each cell type individually. Schelk *et al.* [29] emphasized the importance of creating reference gene expression profiles tailored to single-cell RNA-seq (scRNA-seq) data. Cobos *et al.* [30] conducted a comprehensive and quantitative evaluation of the effects of data conversion, scaling/normalization, marker selection, cell type composition, and method selection on deconvolution results. Jin *et al.* [31] systematically studied the impact of various biological and technical factors. Decamps *et al.* [32] demonstrated that accounting for biological and technical heterogeneities in constructing cell-type reference profiles can enhance the accuracy of deconvolution. In this manuscript, we refer to reference gene expression profiles and cell-type reference profiles—typically represented as a reference matrix of molecular profiles for distinct cell types—as "expression templates." However, previous studies have not fully addressed the variability in optimal deconvolution methods for different cell types within specific datasets, nor have they developed personalized strategies tailored to each cell type. These methodological gaps highlight the need for continued innovation in this area.

Our work is driven by the scientific pursuit of understanding complex biological systems. Building on insights from studies like scCODE [33], which highlighted commonalities in identifying differentially expressed genes across successful methodologies, we present DECEPTICON, a novel deconvolution strategy inspired by a variation of the Anna Karenina principle (AKP). The AKP, popularized by Tolstoy's Anna Karenina, suggests that all successful systems share key traits, while failures arise in diverse ways. This principle has been widely applied across biological contexts to explain how success is marked by consistency, whereas failure exhibits variability. For example, in plant microbiota research, AKP demonstrates that healthy microbiomes exhibit a similar composition, while diseased microbiomes show greater diversity due to disruptions caused by external stressors [34]. In deep learning, AKP distinguishes successful neural network models, which exhibit consistent feature representations, from failed models with higher variability [35]. Similarly, in animal microbiomes, AKP has been instrumental in explaining how healthy microbiomes exhibit compositional consistency, while dysbiosis caused by stress results in greater diversity, often contributing to host health decline [36]. In human microbiome-associated diseases, AKP highlights that healthy microbiomes are consistent across individuals, whereas diseased microbiomes display significant heterogeneity, with the AKP effect observed in ~50% of cases [37]. These applications underscore AKP's value in understanding stability and dysregulation in biological systems. Our approach extends this principle to deconvolution methods, focusing on shared success traits, particularly in the absence of known ground truths in RNA-seq datasets.

Building upon these ideas, we developed DECEPTICON, an innovative approach that synthesizes the strengths of existing deconvolution methods, aimed at addressing these specific challenges. By employing a correlation-based strategy and leveraging scRNA-seq data for dynamic expression template generation, DECEPTICON aims to refine the prediction of cellular proportions within tumor environments. This approach not only improves the precision of deconvolution but also introduces a level of adaptability not previously achievable.

Initial evaluations reveal that DECEPTICON markedly surpasses current methods, as evidenced by its comparison with the top outcomes from 50 analysis strategies—achieving a 23.9% increase in correlation and a 73.5% reduction in root mean square error (RMSE). Applied to The Cancer Genome Atlas (TCGA) datasets for various cancers—including 309 cases of cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), 1228 cases of breast-invasive carcinoma (BRCA), and 553 cases of lung adenocarcinoma (LUAD)—DECEPTICON accurately differentiated between patients with varying prognoses more effectively than existing methods, affirming its robust capability for prognostic differentiation across diverse cancer types.

Results

In this study, we explored a diverse array of methodologies for the deconvolution of bulk RNA-seq data, as outlined in Supplementary Table S1. Our focus was on those methods capable of integrating variable expression templates [18, 20, 38, 39], defined as cell-type-specific gene expression signatures, across various cell types and datasets. Among the various deconvolution methods available, BATMAN [22, 40], originally developed to deconvolute metabolite concentrations from nuclear magnetic resonance (NMR) spectral data, has been applied to deconvolute bulk RNA-seq data for the first time. BATMAN utilizes Bayesian models with a Markov chain Monte Carlo algorithm to automatically quantify complex metabolite mixtures. However, the mathematical infrastructure of BATMAN also allows it to perform RNA-seq deconvolution by replacing NMR data with bulk data and "PureSpectra Template" with a cell-type RNA expression template. It has shown promising performance in predicting the proportion of cell types in bulk samples. (more details can be found in Methods). Five expression templates (ciber [20], TRef [21], BRef [21], quan [23], immu [41]), and the 10 deconvolution methods were combined into 50 analysis strategies. We evaluated each analysis strategy in three aspects: accuracy, sensitivity, and spillover (the incorrect assignment of one cell type to others is called spillover) [28] (Fig. 1a) and compared the performance of these analysis strategies in various application scenarios. Furthermore, we proposed DECEPTICON, which can robustly predict the abundance of cell types from bulk samples, and evaluated its performance on both simulated and real data (Fig. 1b).

Evaluation of simulated bulk RNA-seq data

We first used a single-cell RNA-seq dataset [29] from three different human tissues containing > 11 000 cancer cells, stromal cells, and immune cells to create a simulated bulk RNA-seq dataset with 100 samples and 17 753 genes. Each sample contained both immune cells and tumor cells in different proportions (see Methods). We applied the 51 analysis strategies

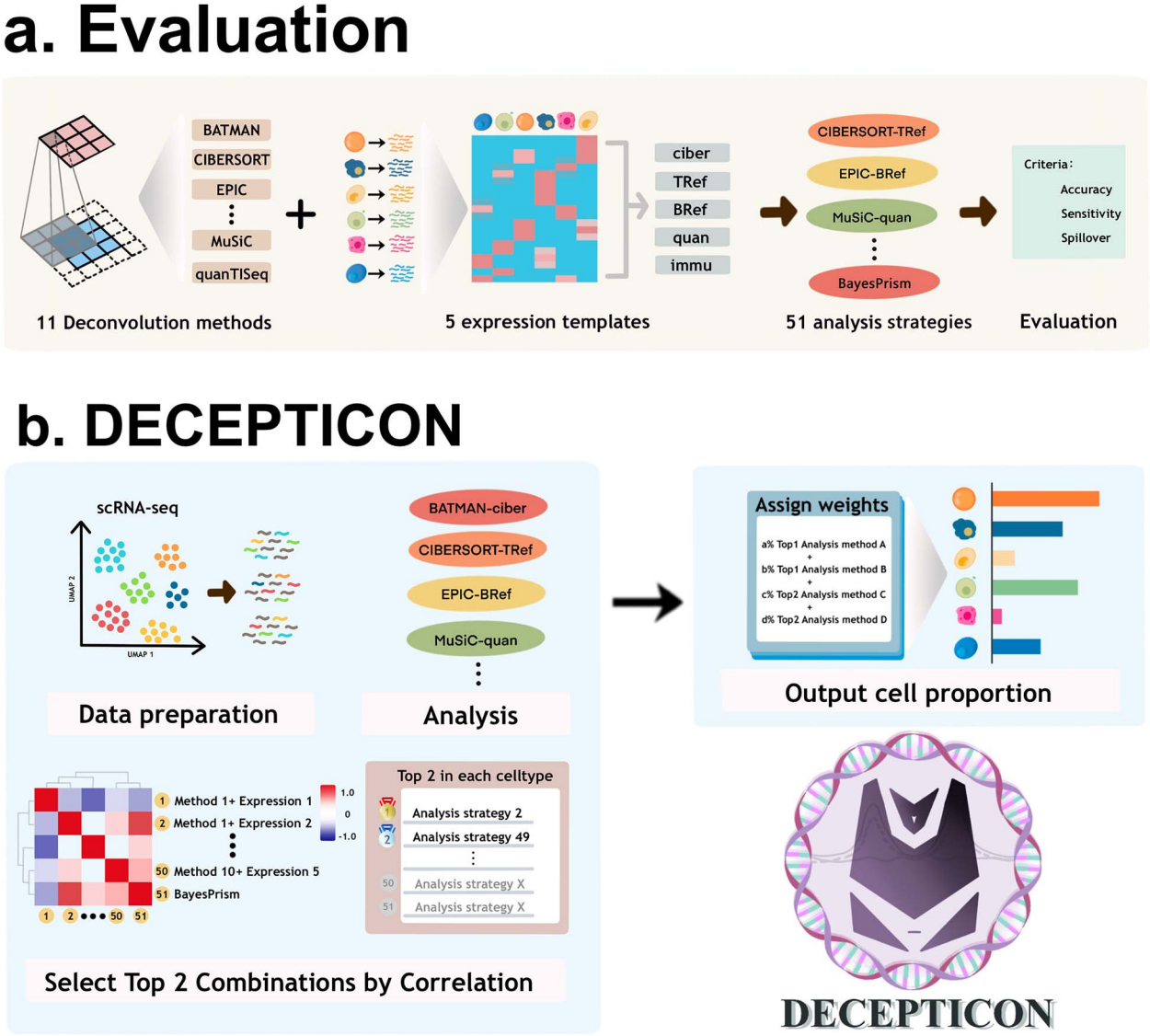

Figure 1. Evaluation of analysis strategies and schematic of DECEPTICON. (a) Evaluation of the performance of different analysis strategies. (b) Schematic of DECEPTICON.

to these samples and compared the estimates of each strategy to the ground truth. Figure 2a shows the results of five analysis strategies (the results of the remaining analysis strategies are shown in Supplementary Figs. S1–S3). As indicated in the figure legend, grids with no data points represent cases where the reference gene expression profile was not available for the cell type. The diagonal slashes (as indicated by different line colors) in these grids are markers indicating the absence of data points for these specific cell types, rather than trend lines. As such, the x and y axes are omitted for these specific plots. Most of the analytical strategies achieved high correlations for B cells (above 0.81, $P < .05$), $CD8^+$ T cells (above 0.82, $P < .05$), and natural killer (NK) cells (above 0.75, $P < .05$) (Fig. 2a). However, regulatory T cells (Tregs), $CD4^+$ T cells, and dendritic cells (DCs) did not show high correlations. For individual cell types, BATMAN performed stably well on all immune cells (Fig. 2c). When comparing the performance of different expression templates, quan and ciber are superior to other expression templates in accuracy for most cell types. In particular, quan is excellent in predicting the abundance of Tregs. Generally, BATMAN and EPIC achieved overall higher accuracy and better stability than the other methods (Fig. 2b). We noticed that some deconvolution methods with their own expression templates (e.g. EPIC (deconvolution method)-TRef (expression template): $r = 0.74$, $P < .05$, quan-quan: $r = 0.61$, $P < .05$) performed better than when integrated with other expression templates (e.g. EPIC-ciber: $r = 0.35$, $P < .05$, quan-immu: $r = 0.32$, $P < .05$) (Fig. 2b). In contrast, CIBERSORT and its expression template (ciber-ciber: $r = 0.42$, $P < .05$) obtained poorer performance than the analysis strategy of CIBERSORT and TRef (ciber-TRef: $r = 0.54$, $P < .05$). This suggests that the combination of deconvolution methods and expression templates requires more careful selection for the given datasets and the cell types of interest. We also used the coefficient of determination (R^2) and RMSE to further evaluate the prediction accuracy of each analysis strategy. A higher correlation and R^2 value, along with a lower RMSE value, indicate better prediction accuracy for the analysis strategy. Overall, EPIC, quanTIseq, and CIBERSORT-abs demonstrated higher accuracy and better stability than the other methods. The average R^2 and RMSE values for EPIC, quanTIseq, and CIBERSORT-abs were (0.51, 0.08), (0.47, 0.12),

and (0.44, 0.16), respectively, and their R^2 values also ranked among the top, generally outperforming other methods (Fig. 2b, Supplementary Fig. S4a, Supplementary Fig. S17a).

The recent development of MuSiC2 [42], BayesPrism [43], etc., can dynamically infer cell type expression templates from single-cell expression data. MuSiC2 requires bulk RNA-seq data generated from samples from multiple clinical conditions. BayesPrism is a Bayesian method that uses patient-derived scRNA-seq as prior information for deconvolution. As BayesPrism has an input data format that is more compatible with other analysis strategies, we also incorporated a comparison with BayesPrism in this study. Specifically, compared with the top three analysis strategies (MCP-quan, EPIC-TRef, and batman-ciber) with fixed expression templates that had the highest average correlation among the 50 analysis strategies, BayesPrism outperformed them with a smaller RMSE (0.019) (Supplementary Fig. S4) and a higher average correlation ($r = 0.86$, $P < .05$) (Supplementary Fig. S5a). In particular, BayesPrism showed higher prediction accuracy for macrophages/monocytes, Tregs, and NK cells (Supplementary Fig. S5a).

Accuracy for predicted cell subtype abundance

In the above section, we evaluated the performance of deconvolution methods in estimating the proportion of major cell types. In this section, we systematically evaluated the capability of different analysis strategies in deconvolving cell subtypes. We created 100 simulated bulk RNA-seq samples of cell subtypes with various signal-to-noise ratios (SNRs) using the expression template ciber from CIBERSORT. Each sample contains five types of cell subtypes (see Methods). We applied the analysis strategies to these samples and compared the predicted abundances of the cell subtypes with the ground truth. We observed that most methods performed well in predicting DC cell subtypes and macrophage subtypes and were less affected by noise. However, when predicting B cell subtypes, most methods were significantly impacted by noise (Fig. 3a, Supplementary Fig. S5b). When considering all subtypes of the five cell types together, quanTIseq, EPIC, and SCDC stood out as outstanding performers (Fig. 3b, Supplementary Fig. S5c). Regardless of how the noise varied, these methods consistently maintained a relatively high correlation compared to other methods. This demonstrates that quanTIseq, EPIC, and SCDC have distinct advantages in predicting cell subtypes.

Sensitivity for low-abundance cell types

To evaluate the sensitivity of different analysis strategies, we applied two metrics similar to those presented in [28], i.e. the Minimum Detection Fraction (MDF) and the Background Fraction (BF) of each cell type (see Methods for definitions). The simulated datasets consist of background cells and a gradient proportion of the target cell type. The MDF value represents the minimum cell fraction that each analysis strategy can reliably identify cell types. The BF value represents the cell fraction predicted by each analysis strategy when a certain cell type is actually absent in bulk samples. Lower values of both metrics indicate higher sensitivity of the analysis strategy. We observed that the MDF values for most analysis strategies were low for B cells, which indicated that the majority of analysis strategies were sensitive enough to predict B cells. Regarding $CD4^+$ T cells, a majority of analysis strategies had high BF values, while certain analysis strategies exhibited high MDF values (such as batman-immu, scdc-quan, and music-BRef) (Supplementary Figs S7–S10). When comparing different deconvolution methods, we found that in most cases, the MDF values obtained by quanTIseq and EPIC were smaller than those obtained by the other methods. In addition, MuSiC and SCDC have low BF values for $CD4^+$ T cells, $CD8^+$ T cells, NK cells, and macrophages/monocytes at the cost of high MDF values (Fig. 4, Supplementary Fig. S7–S10). When comparing different expression templates, we observed that in most cases, the MDF values obtained by the expression template quan coupled with various deconvolution methods tended to be small (Fig. 4, Supplementary Fig. S7–S10).

Spillover for different cell types

The BF analysis of each approach reveals that its score may be detected even in the absence of the corresponding cell type in the bulk sample. Sturm *et al.* [28] suggested that this may be due to nonspecific signature genes of cell types. Therefore, in this section, we compared the spillover of each analysis strategy to investigate which cell types were incorrectly assigned as another. We used simulated bulk samples containing only one cell type of interest to assess the spillover of each analysis strategy. We recorded the percentage of cell types that were incorrectly assigned by each analysis strategy (the error rate). Considering the performance on different cell types, different analysis strategies performed diversely. For example, EPIC-TRef can predict pure B cell and $CD8^+$ T cell samples more accurately, with error ratios as low as 0.16 and 0.06 for these two cell types. However, it predicted pure $CD4^+$ T cell samples with an error ratio as high as 0.41 (Fig. 5a). Ciber-TRef performed well in predicting pure $CD4^+$ T cell samples with an error ratio as low as 0.15, although it had a poor average error ratio (0.73) and was not good at predicting pure B cell samples and pure $CD8^+$ T cell samples (error ratios 0.52 and 0.98, respectively) (Fig. 5b). These results showed that integrating the advantages of different analysis strategies on various cell types may help improve overall performance. Notably, since the expression template TRef does not contain Tregs and DCs, the corresponding results were incorrectly classified as other cell types. The overall error rate was reduced by adding the expression templates of Tregs and DCs, such as EPIC-quan (0.37) and ciber-ciber (0.53) (Supplementary Fig. S6b).

Deconvolution performances vary with different datasets

In this section, we further applied the analysis strategies to five types of specimens: peripheral blood mononuclear cells (PBMCs) [44], whole blood [45], head and neck squamous cell carcinoma (HNSCC) [45], melanoma [21], and ovarian cancer [29]. All five datasets had "gold-standard" cell type composition estimates obtained by FACS or flow cytometry. In our analysis, we did not exclude any cell types from the dataset or reference matrix. All cell types defined in the expression templates were estimated for cell composition. However, to facilitate meaningful comparisons between different templates, we only presented the results for the cell types that were shared among the expression templates. (The full results are illustrated in Supplementary Tables S3–S7.) Considering the performance of different expression templates, we found that the expression template quan performed well on all types of datasets, while other expression templates performed differently. For example, ciber performed excellently on whole blood and HNSCC datasets but less well on ovarian cancer and melanoma; immu performed second best on the whole blood, melanoma, and HNSCC datasets but not on the ovarian cancer dataset (Fig. 5c). Considering the types of datasets, across the five types of datasets, the top five best-performing strategies vary widely. Furthermore, there is no single strategy that fits all data types (Fig. 5d). For instance, MCP-quan achieved high average correlation, high R^2 value, and low RMSE in the melanoma

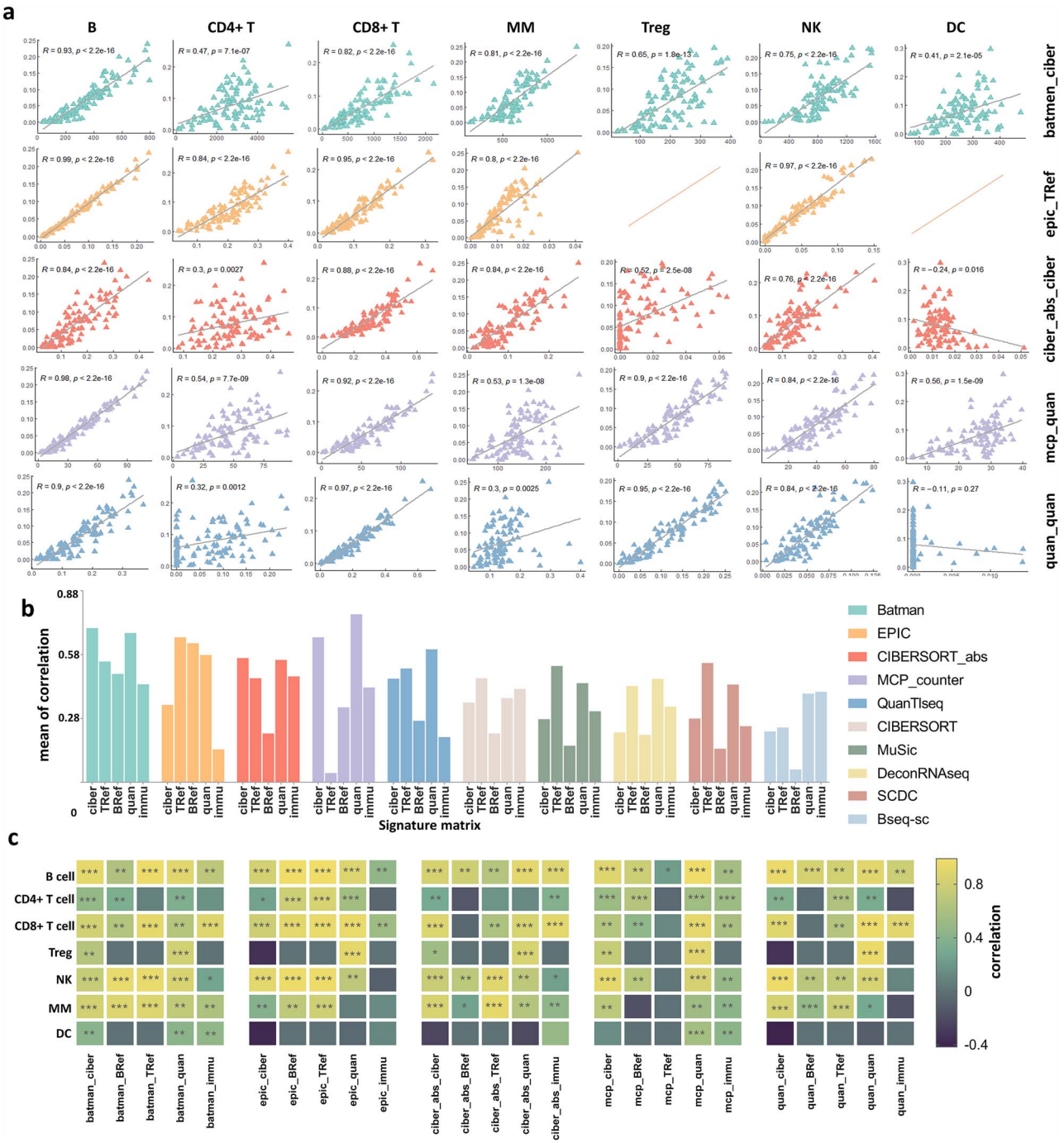

Figure 2. Comparing the accuracy of 50 analysis strategies capable of predicting the proportion of major immune cell types on simulated data. (a) Scatterplot of known cell proportions and estimated cell proportions on simulated data. Additional plots are shown in Supplementary Fig. 1–3. r indicates Pearson's correlation coefficient, and p indicates the *p*-value. Grids with no data points indicate that the reference gene expression profile does not provide a signature for the corresponding cell type. (b) Histogram of the mean of the correlations of the estimated and known fractions for the 50 analysis strategies. Each color represents a method and corresponds to Fig. 2a and Supplementary Fig. 2. (c) Heatmap of the correlation between the predicted and known fractions of 7 immune cell types in 25 analysis strategies.

dataset, but exhibited low correlation, low R^2 value, and/or high RMSE in the other four datasets. Additionally, music-quan, which achieved a higher average correlation, higher R^2 value, and lower RMSE in the PBMC dataset, performed poorly in the remaining datasets (Fig. 5d, Supplementary Figs S11a and S17b). Our analysis showed that individual strategies generally fail to achieve high correlation, high R^2 value, and low RMSE simultaneously across datasets. This suggests that we need to adjust deconvolution methods for different types of datasets. It is worth noting that the average correlations are low in all of the analysis strategies in the ovarian cancer dataset. This may be due to the small abundance of most of the cell types (Supplementary Table S2), and certain cell types were poorly deconvoluted in various strategies (Fig. 6b), ultimately impacting their overall performances. However, not all cell types exhibit poor predictive performance, and specific subsets demonstrate promising results. For instance,

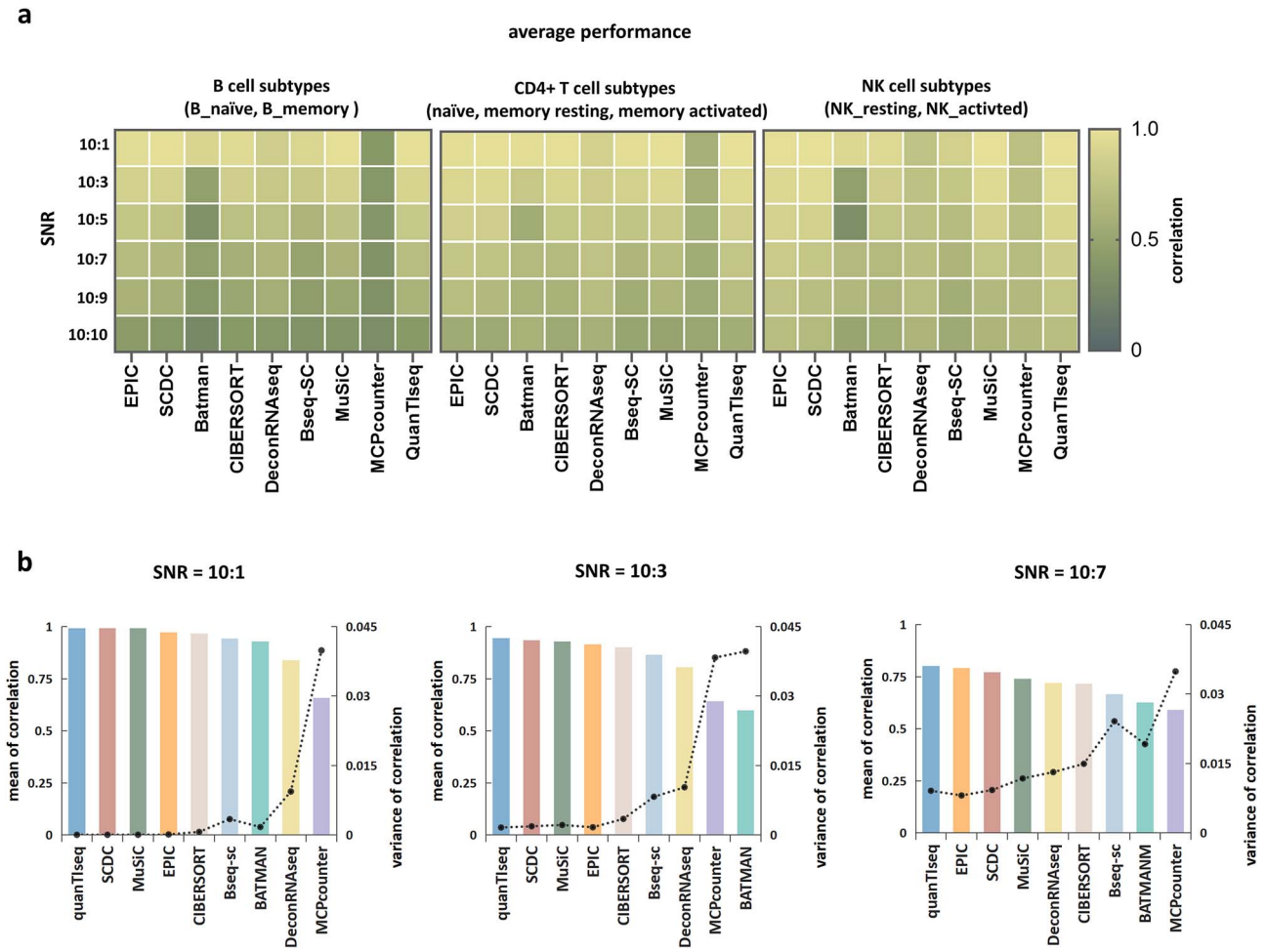

Figure 3. Assessing the predictive accuracy of analysis strategies for cell subtypes on simulated data. (a) Heatmap of the correlation between predicted values and known proportions for B cell subtypes, CD4+ T cell subtypes and NK cell subtypes. (b) Histogram of the mean of the correlation between the estimated and known fractions of the analysis strategies at signal-to-noise ratios of 0.1, 0.3, and 0.7. Each color represents a method and corresponds to Figure 1b. The left vertical coordinate represents the mean of the correlations for the five cell subtypes; and the right vertical coordinate represents the variance of the correlations.

quan-quan demonstrates good predictive capabilities for B cells and macrophages/monocytes, whereas its results for CD4+ T cells are less favorable (Fig. 6b).

DECEPTICON: optimizing deconvolution with a Tolstoyan perspective

Overall, by systematically evaluating the performance of different deconvolution methods, we concluded that deconvolution strategies must be tailored for each individual cell type in each specific dataset, which was not noticed in previous studies. Given the dearth of automated approaches capable of discerning optimal deconvolution methods for particular cell types within a given dataset, we devised DECEPTICON (DEconvolution for CEll Proportion esTImation by COrrelatioN). DECEPTICON, inspired by the principle that "All happy families are alike," applies a correlation-based strategy to finely tune deconvolution to specific cell types and datasets, achieving unmatched accuracy. The Pearson correlation coefficient (ρ) serves as our metric for evaluating the similarity of the outputs of deconvolution methods, drawing parallels to the concept of "alikeness" as illustrated in Tolstoy's narrative. DECEPTICON leverages mathematical evidence suggesting that successful deconvolution results, similar to Tolstoy's notion of happy families, tend to have higher correlation coefficients among themselves, while less successful ones display lower coefficients (for more details, refer to the supplementary methods section). This principle underlies our selection criterion, which asserts that the most effective analysis strategies for deconvolution demonstrate significant concordance (Supplementary Figs S12–S14).

For example, in the simulated data, strategies from the top two highest pairwise correlations for CD4+ T cells (MCP-ciber, EPIC-TRef, MCP-quan, and EPIC-BRef) were selected and assigned equal weights (0.25) to obtain the final estimated value. Similarly, for B cells, EPIC-BRef, EPIC-TRef, MCP-quan, and EPIC-BRef were selected, with the weight of EPIC-BRef doubled due to its repeated appearance.

We also conducted sensitivity analyses using different weighting schemes based on method correlation rankings. However, we found that the equal weighting strategy produced more consistent and reliable results across various datasets. Therefore, we adopted the uniform weight approach as the final weighting scheme.

To evaluate the performance of DECEPTICON, we applied it to the same simulated bulk RNA-seq dataset described in the "Evaluation of Simulated Bulk RNA-seq Data" section, which was constructed from single-cell RNA-seq data of three different human tissues and contains over 11 000 cancer, stromal, and immune cells in varying proportions (see Methods). Figure 6a

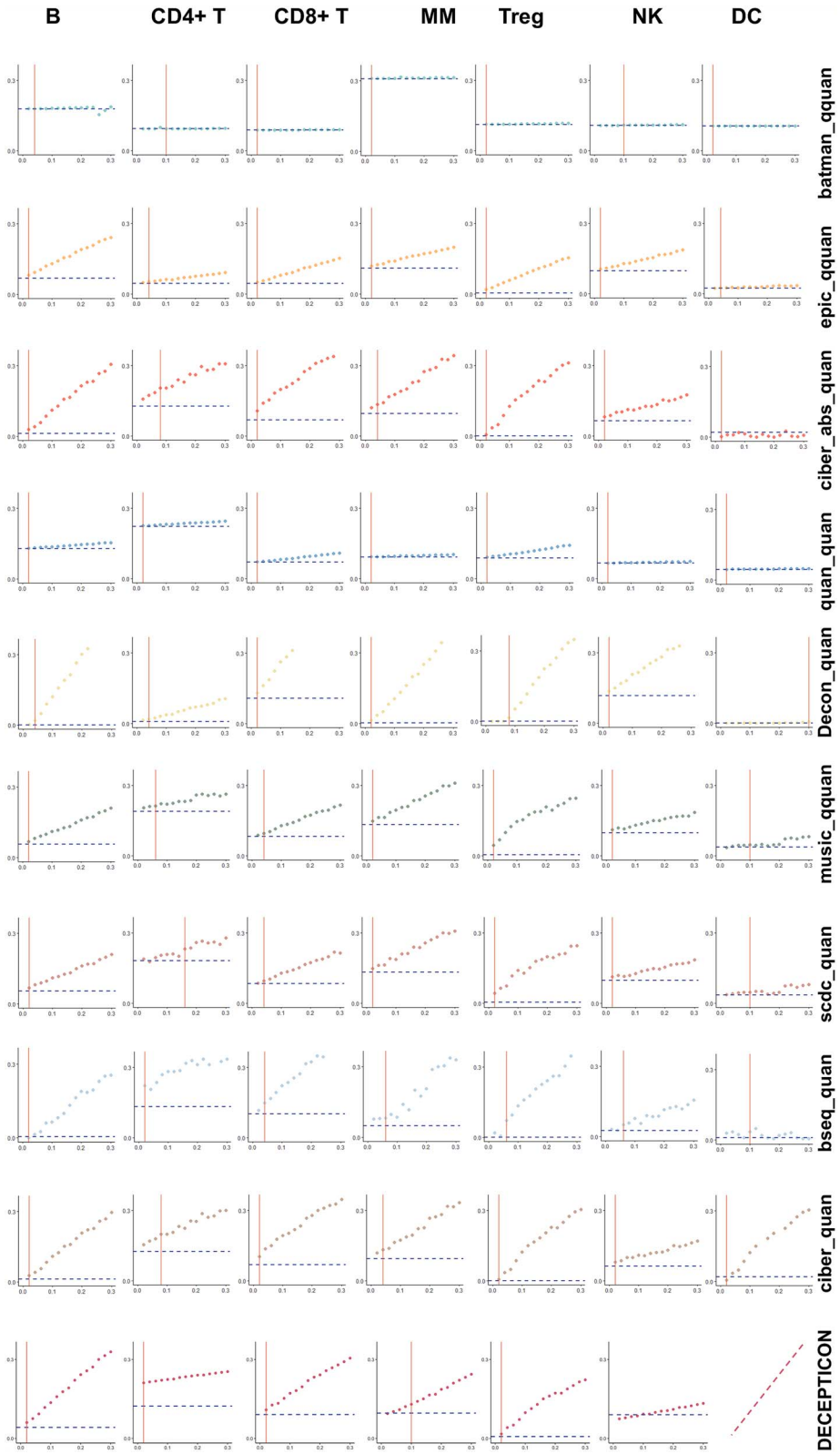

Figure 4. The sensitivity of the analysis strategies shows differences in the simulated data. Scatter plots of the minimum detection fractions and background fractions for the six analysis strategies. The horizontal coordinate represents the proportion of cell types in the sample. The vertical coordinates represent the mean predicted values. Each point represents the predicted mean of 15 simulated samples. The red line indicates the minimum detection fraction. The blue line indicates the background score.

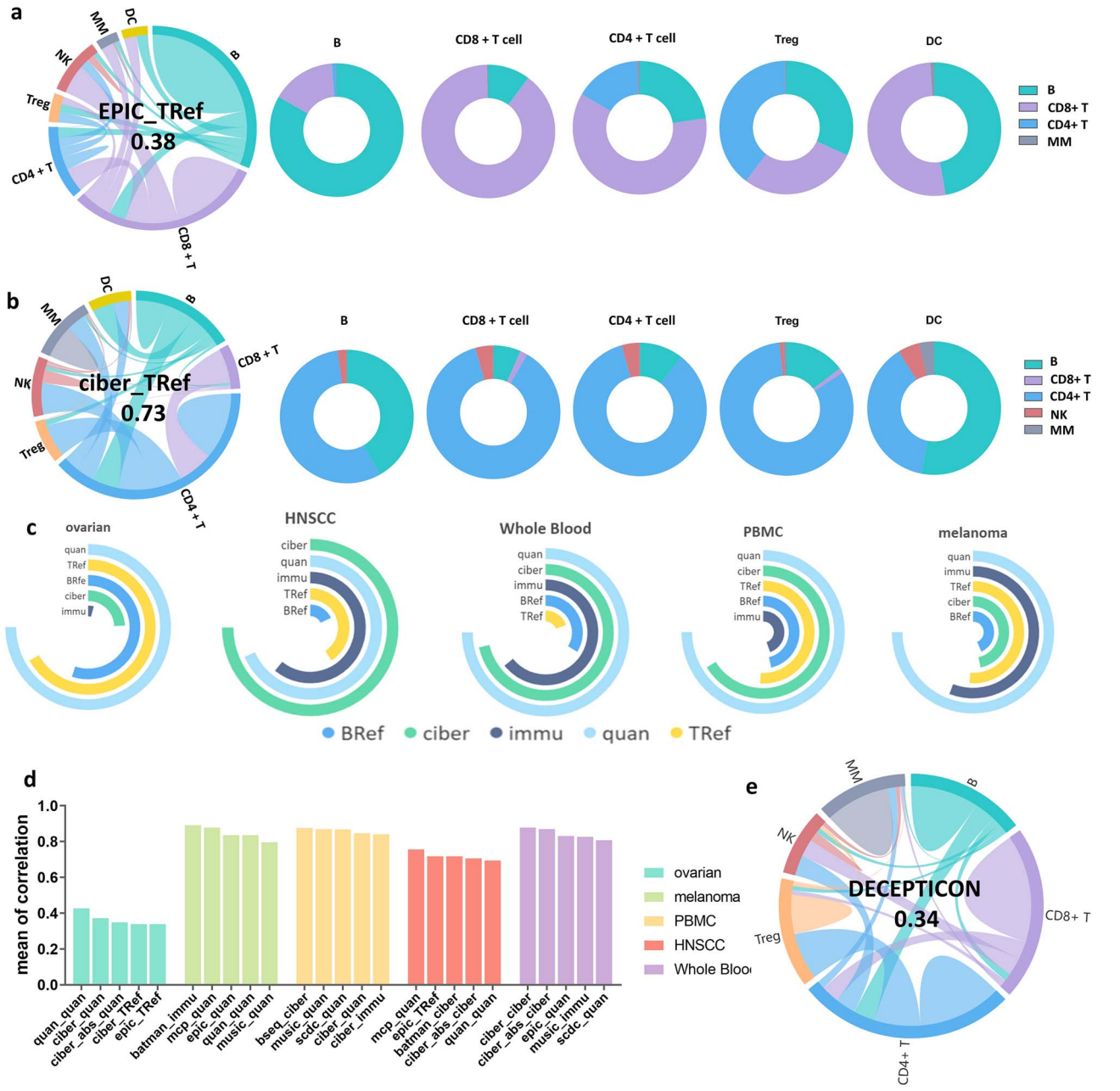

Figure 5. Comparison of the spillover and the differences in the five datasets of the analysis strategies. (a) String plots of spillover analysis of EPIC-TRef. Pie charts of predicted cell composition in pure B-cell, $CD8^+$ T cell, and $CD4^+$ T cell samples. (b) String plots of spillover analysis of Ciber-TRef. Pie charts of predicted cell composition in pure B cell, $CD8^+$ T cell, and $CD4^+$ T cell samples. The number in the middle indicates the average percentage of incorrectly predicted cell types. (c) Radial bar chart of the average correlation of the reference gene expression profiles across the 5 datasets. The length of the arcs indicates the relative average correlation. (d) Bar plot of the top 5 performing analysis strategies in each of the 5 datasets. (e) String plots of spillover analysis of DECEPTICON. The number in the middle indicates the average percentage of incorrectly predicted cell types.

evaluates the accuracy of DECEPTICON using the same simulated datasets as Fig. 2b, highlighting the superior performance of DECEPTICON. Compared with other analysis strategies, DECEPTICON achieves the highest average correlation ($r = 0.90$, $P < .05$) and R^2 value (0.88) (Supplementary Fig. S17a), along with the lowest variance (0.008) (Fig. 6a) and RMSE (0.008) (Supplementary Fig. S4), demonstrating its high accuracy and stability. Furthermore, Fig. 4, shows that that DECEPTICON demonstrated higher sensitivity in detecting B cells and Treg cells. Additionally, Fig. 5e revealed that DECEPTICON exhibited a lower spillover rate compared to other analytical strategies. These results further highlight the robust performance of DECEPTICON.

We further validated DECEPTICON on three tumor datasets: HNSCC [40], ovarian cancer [29], and melanoma [45]. The ground truth data for these datasets was used exclusively for performance evaluation without tuning DECEPTICON's outputs. We compared DECEPTICON's estimated cell-type proportions with the ground truth to evaluate its accuracy. The cell abundance estimates obtained by DECEPTICON and the gold standard were highly correlated for most immune cell types. DECEPTICON outperformed other analysis strategies in terms of correlation, $R^{2,}$ and RMSE across different datasets. Specifically, DECEPTICON exhibited higher average correlations ($r = 0.91$, $r = 0.74$, and $r = 0.89$ for HNSCC, ovarian cancer, and melanoma, respectively, with all

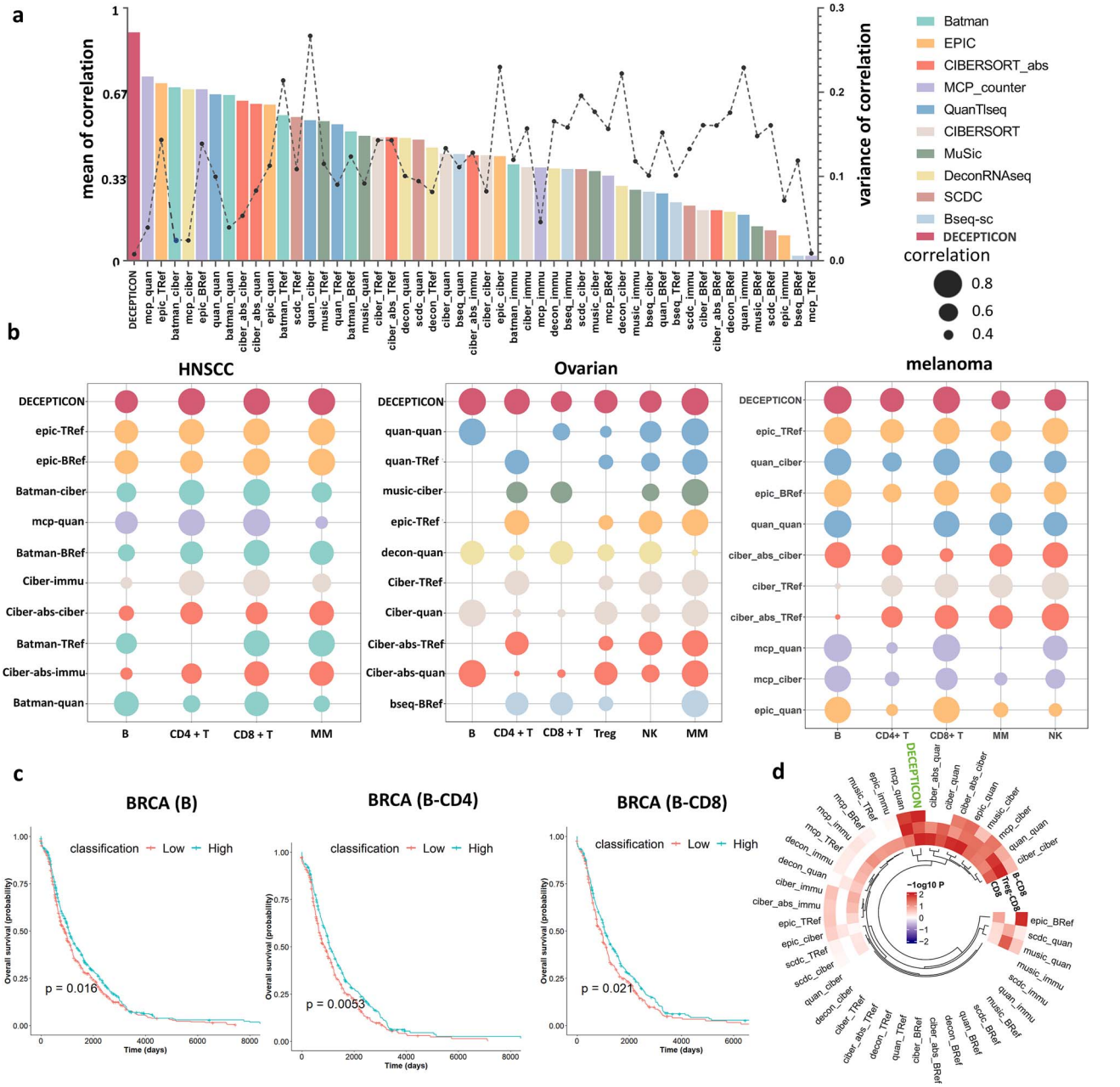

Figure 6. Validation of DECEPTICON on simulated and real datasets. Prognostic value of DECEPTICON-based immune scores and T and B cell scores for solid cancers. (a) Bar plot comparing DECEPTICON to 50 other analysis strategies on a simulated data set. (b) Bubble plots of the top 10 performing analytics and DECEPTICON comparisons for the three-tumor data. The size of the bubbles indicates the correlation between the predicted value of the analysis method and the "gold standard" ratio; the larger the bubble is, the higher the correlation. (c) Kaplan-Meier plot showing the overall survival rate based on B scores, B-CD4 scores, and B-CD8 scores calculated by DECEPTICON for patients with endometrial cancer. The red line indicates patients with scores below the median, and the blue line indicates patients with scores above the median. (d) Cyclic heat map based on negative log10p values for the three scores calculated by DECEPTICON and other analysis strategies. Each circle represents one score; each column represents one analysis method.

corresponding $P < .05$), and the R^2 values (0.83, 0.60, and 0.71 for HNSCC, ovarian cancer, and melanoma, respectively) across three distinct datasets compared to other strategies (Fig. 6b, Supplementary Fig. S17c). The average RMSE values (0.059, 0.079, and 0.06 for HNSCC, ovarian cancer, and melanoma, respectively) were also lower than those of the majority of analysis strategies (Supplementary Fig. S11b). In contrast, individual analysis strategies showed variable performance across datasets. For instance, EPIC_TRef had a high average correlation of 0.89 in HNSCC and melanoma but only 0.34 in the ovarian dataset. Quan-quan accurately predicted certain cell types in the ovarian dataset, such as B cells ($r = 0.98$, $P < .05$) and macrophages/monocytes ($r = 0.99$, $P < .05$), but performed poorly in predicting CD4$^+$ T cells (Fig. 6b). By integrating the strategies that produced good results for each cell type, DECEPTICON improved the accuracy and overall stability across datasets (Fig. 6b, Supplementary Fig. S11b). In conclusion, DECEPTICON is broadly applicable and consistently delivers relatively optimal results.

In addition, incorporating BayesPrism into the framework of DECEPTICON further improves the deconvolution performance

(Supplementary Fig. S15a). This demonstrates the ability of DECEPTICON to continuously improve its performance by incorporating appropriate deconvolution methods or expression templates. It is worth noting that DECEPTICON including BayesPrism outperformed BayesPrism alone.

Refining cell type analysis with DECEPTICON via scRNA-seq data

DECEPTICON leverages scRNA-seq datasets to infer cell type fractions, addressing the limitations of existing cell type expression templates, which are predominantly focused on immune cells and often lack gene expression data for cell subgroups. DECEPTICON allows users to input either custom expression templates or scRNA-seq data, facilitating the accurate deconvolution and estimation of cell type compositions across large datasets. When scRNA-seq data are provided, DECEPTICON incorporates BayesPrism into the analysis to enhance the prediction of cell types. The final composition of each cell type is determined through correlation analysis across various strategies. In datasets such as HNSCC, melanoma, and ovarian cancer ascites, DECEPTICON demonstrated comparable performance between the versions utilizing single-cell data-derived templates and those relying on inherent templates alone. Particularly in the PBMC dataset, the single-cell data-informed DECEPTICON showed enhanced performance, likely influenced by the quality of the selected single-cell dataset (Supplementary Fig. S11c). Overall, DECEPTICON demonstrates proficiency in dynamically inferring cell-type expression templates from single-cell datasets, significantly enhancing the accuracy of cell composition inference when appropriate datasets are selected.

DECEPTICON-based prediction scores for cancer prognosis

The proportion characteristics of lymphocytes have been shown to be superior to prognostic markers for colorectal cancer (CRC) [38]. Previous studies also showed that T-B cells (TB score) are a valid predictor of disease-free survival and overall survival in metastatic CRC [39]. Here, we defined prediction scores based on the proportion of individual immune cells and tested their relationship with solid tumor prognosis (see Methods). Survival analysis using the calculated cell proportion scores showed that the DECEPTICON-based prediction scores can discriminate good/poor-prognosis patients in CESC (Supplementary Fig. S15b–d). When the same prediction scores were calculated using other analysis strategies, we found that the prediction results highly fluctuated (Supplementary Fig. S15e). Similarly, when applied to 1228 BRCA samples from TCGA, DECEPTICON still effectively distinguishes patients with good and poor prognoses (Fig. 6c). Although the survival curves for the two patient groups stratified by DECEPTICON scores show relatively small differences, statistical analysis revealed a significant survival difference ($P < .05$). The small gap between the curves may be attributed to the overlap in survival distributions between the two groups, which reduces the visibility of the difference. Nevertheless, the statistically significant P-value suggests that DECEPTICON-based prediction score has a meaningful prognostic value. Furthermore, DECEPTICON-based prediction scores yielded more consistent and stable results in distinguishing patient prognosis compared to other analytical methods (Fig. 6d). In contrast, a single method did not consistently produce reliable results across different datasets. For instance, MCP-quan performed well in the CESC dataset, particularly for Treg-CD8$^+$ T cells, but showed comparatively poorer performance in the BRCA datasets.

Discussion

Quantifying tumor-infiltrating immune cells from bulk RNA-seq data is essential for understanding the tumor microenvironment. While numerous deconvolution methods exist, their capabilities and limitations vary, making it crucial to identify the best strategy for each cell type in specific datasets. Our study assessed 51 different analysis strategies across five distinct clinical datasets to identify the most suitable approaches for diverse cell types.

Our findings demonstrated that the performance of combining deconvolution methods and different expression templates varied considerably, such as MCP-counter (the average correlation of MCP-quan: $r = 0.75$, $P < .05$, MCP-TRef: $r = 0.47$, $P < .05$) (Fig. 2b). Moreover, we observed that different specimen types and cell types require specific optimal method(s). The analysis strategy that achieved the best performance varied among the five data types (BATMAN-immu $r = 0.89$, in melanoma, bseq-ciber 0.88 in PBMCs, mcp-quan 0.76 in HNSCC, and ciber-ciber 0.88 in whole blood, with all corresponding $P < .05$) (Fig. 5d). Considering the differences in results on cell types, the analysis strategy with the best performance in B cells of ovarian cancer was ciber-quan ($r = 0.99$, $P < .05$) and in CD4$^+$ T cells was music-BRef ($r = 0.86$, $P < .05$). Due to the low abundance of DCs in single-cell data sets and the ambiguous heterogeneity across DC subtypes [46, 47], most analysis approaches are ineffective in evaluating DCs (Supplementary Fig. S1-S3). Sturm *et al.* [28] previously observed this phenomenon and neglected the deconvolution of DC subtypes.

Notably, we applied BATMAN to deconvolute RNA-seq data for the first time. The combination of BATMAN with the five expression templates performs outstandingly in accuracy, which indicates that it is not susceptible to the choice of expression templates and has better stability.

Our study also presents DECEPTICON, a correlation-based approach inspired by the AKP, which posits that successful outcomes often share common traits while failures are more variable. DECEPTICON improves accuracy across diverse cell types by leveraging shared patterns of success among methods, even in the absence of ground truth data. Notably, DECEPTICON's performance in analyzing scRNA-seq data in cancers like HNSCC and melanoma highlights its adaptability and potential impact on computational biology. Previous studies demonstrated that analysis results can vary substantially depending on the deconvolution method and expression template chosen [48]. Our findings further emphasize that optimal strategies may differ by cell type, and DECEPTICON's self-adapting approach identifies the most suitable methods for each cell type in the dataset.

We validated DECEPTICON on both simulated and real data. The cell proportions estimated by DECEPTICON were more consistent with the ground truth than those obtained by the individual analysis strategies. Moreover, DECEPTICON also demonstrated its capability to more reliably discriminate tumor patients with good/poor prognoses. Notably, most deconvolution methods tested feature their own normalization step when outputting results (sum equals 1) to ensure that they are unaffected by the expression templates. For the two exceptions, BATMAN and MAPcounter, we normalized them with the same standard in DECEPTICON to ensure comparability with other methodologies. Furthermore, methods like quanTIseq and EPIC account for the unknown cell type fraction in their outputs. In contrast, DECEPTICON typically reports immune cell types derived from its built-in expression templates. When single-cell data is incorporated, DECEPTICON also reports the identified cell types from that dataset. Consequently, DECEPTICON's output does not include an unknown cell type fraction.

These results demonstrate the robustness of DECEPTICON in cancer prognosis and highlight its broad applicability. The key distinction between DECEPTICON and both SQUID [48] and granulator [49] lies in their application contexts. SQUID was specifically designed to address the coexistence of scRNA-seq and snRNA-seq datasets. While it excels in this narrow scope, its applicability is limited when only bulk RNA-seq data are available, restricting its use in large-scale tumor studies. In contrast, granulator focuses on benchmarking and comparing existing deconvolution methods, offering valuable insights for method selection. However, it does not provide a comprehensive solution or an integrated approach for consistent analysis, which limits its practical use across varied datasets.

In contrast to these methods, our evaluation of analytical strategies is more comprehensive. Specifically, while granulator evaluates different deconvolution methods solely based on accuracy using four metrics (pcc, ccc, adj.r^2, rmse), we assess strategies from three perspectives: accuracy, sensitivity, and spillover, providing a more thorough analytical framework. Additionally, we propose DECEPTICON, which employs a correlation-based strategy that dynamically integrates multiple top-performing deconvolution methods. This approach is particularly effective for bulk RNA-seq datasets, where traditional single-method approaches often struggle with heterogeneous samples. DECEPTICON is applicable not only to large-scale tumor datasets, such as those from TCGA (including breast cancer, CESC, and LUAD), but also to a wide range of transcriptomic data types. By leveraging a correlation mechanism, DECEPTICON significantly improves the accuracy and robustness of deconvolution, resulting in more precise cell-type abundance estimations and enhanced prognostic stratification of cancer patients.

In the context of unknown cell types present in bulk data, both quanTIseq and EPIC account for these unknowns in their outputs. Given that DECEPTICON integrates the quanTIseq and EPIC algorithms, it similarly considers unknown cell types in its calculations. However, as our analysis does not focus on the proportions of these unknown cell types, DECEPTICON's output does not discuss them. The sum of cell type estimates generated by DECEPTICON does not equal one, which restricts within-sample comparisons. However, this limitation does not affect the utility of DECEPTICON, as most deconvolution algorithms are intended for between-sample analysis. For example, in TCGA data that incorporates clinical information, DECEPTICON can predict patient prognosis using either individual cell type estimates or the ratio of estimates from two distinct cell types. The full version of DECEPTICON can be computationally intensive. For example, the runtime for the HNSCC dataset (23 687 genes × 23 samples), with Intel i7-11700 2.5G Hz, 64 GB RAM and 8 cores, was ∼200 min. To address this, we developed a "light" version, reducing the runtime to ∼ 2.35 min with only a slight decrease in performance. Despite this reduction, the light version still outperformed other top analysis strategies (Supplementary Fig. S5a). We will continue to include more methods and expression templates to complement DECEPTICON to further improve performance.

Conclusion

Reflecting a variation of the AKP, our work with DECEPTICON demonstrates its adaptability and efficacy across various datasets, both simulated and experimental. This method excels at identifying optimal outcomes, particularly in scenarios lacking defined ground truths, resonating with a broader AKP. It suggests the potential for pioneering new methodologies in biological research, aiming for enhanced accuracy and deeper insights into complex biological systems.

Materials and methods

Overview of the DECEPTICON calculation process

1. *Input data*

DECEPTICON requires the following input data: bulk RNA-seq data for deconvolution analysis, optionally users may choose to utilize DECEPTICON's built-in templates or upload custom *n* set(s) of expression templates or scRNA-seq data.

2. *Selection of optimal analytical strategy via correlation*

After loading the bulk RNA-seq data, DECEPTICON applies it to 10 analytical strategies, generating 10Xn or 50 (with built-in templates) distinct prediction outcomes. Pearson correlation coefficients are computed between these prediction outcomes, and the two pairs of methods with the highest correlation for each cell type are selected as the optimal analytical strategy.

3. *Weighting and integration of results*

For each cell type, the selected methods are initially assigned equal weights (0.25). If a method is selected multiple times, its weight is proportionally increased. The final estimate for each cell type is derived from the weighted sum of these methods.

4. *Normalization*

For certain deconvolution methods (e.g. MCPcounter and BATMAN), DECEPTICON performs additional normalization to ensure that the sum of the results is equal to 1. This normalization ensures that the outcomes from different methods can be compared on a consistent scale.

5. *Handling of zero values*

When two methods exhibit extremely high correlation and both produce zero estimates for a specific cell type, DECEPTICON excludes the deconvolution result for that cell type. This prevents artificial correlations arising from the simultaneous presence of zero values.

The Minimum Detection Fraction and Background Fraction

We first randomly selected 1500 cells as background cells, and then we added varying proportions of cell types of interest to the sample. The cell proportion increased from 0.02 to 0.3 in increments of 0.02, with the number of cells calculated as 1500∗ratio/(1-ratio). This was repeated five times and resulted in five sets of data with 90 samples (15 cell proportions × 6 cell types). Each of the five datasets was submitted to 50 methods individually. The BF for a certain cell type c was quantified as the predicted value of c in samples lacking c-type cells. The final background score was determined by calculating the average of the five batches of samples. The MDF for cell type c was quantified as the lowest cell proportion that exhibited significant deviation from the BF. In our analysis, we used R (version 4.1.3) [50] to perform the calculations, utilizing the "t.test" function to calculate the corresponding *P*-values. The minimum *P*-value that fell below the threshold of 0.05 within the range of cell proportions (0.02–0.3) was selected as the ultimate MDF.

Calculating prognostic prediction scores

In the study by Mlecnik *et al.* [44], ratios between specific immune cell populations, such as $CD8^+$ cells to B cells and $CD3^+$ cells

to $CD8^+$ cells, were identified as being associated with patient prognosis. Building on this foundation, we expand the analysis to a more comprehensive set of immune cell ratios, including Tregs to $CD8^+$ T cells, B cells to $CD8^+$ T cells, B cells to $CD4^+$ T cells, and $CD8^+$ T cells alone, to assess the accuracy of prognostic prediction scores.

The prediction scores were determined by dichotomizing cell fractions across all patients in the cohort. Specifically, patients were classified into two groups for each immune cell type:

- Group 1 (Score = 0): Cell fraction ≤ median value
- Group 2 (Score = 1): Cell fraction > median value

The prediction scores were used to assess patient prognosis by evaluating their ability to distinguish between individuals with good and poor outcomes.

Key Points

- Comprehensive evaluation: Evaluated 50 analytical strategies by combining 10 deconvolution algorithms with five expression templates, focusing on accuracy, sensitivity, and spillover.
- Identified limitations: Current algorithms lack consistent accuracy across various cell types in heterogeneous datasets without ground truth.
- DECEPTICON strategy: Introduced DECEPTICON, an optimal-results selection approach inspired by the Anna Karenina Principle, dynamically integrating top-performing methods to enhance accuracy.
- Enhanced performance and application: DECEPTICON showed a 23.9% improvement in correlation and a 73.5% reduction in root mean square error. Applied to The Cancer Genome Atlas datasets for breast, cervical, and lung cancer, it improved patient prognosis differentiation.

Acknowledgements

We extend our gratitude to Jiayan Zou for her insightful contributions to the manuscript.

Author contributions

F.D., J.H., X.Z. and J.Z. conceived and designed the algorithm. F.D. developed the software. J.H., X.Z., and M.W. provided data interpretation and biological explanation. F.D., J.H., X.Z. and J.Z. wrote the manuscript, and J.H., X.Z., J.Z., F.D. and M.W. prepared the figures. Y.G., J.W. and L.G. tested the software. Y.C., Y.J., H.Y.T., J.C., W.C., and L.T. provided critical suggestions. The manuscript was approved by all authors.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Funding

This work was supported in part by the National Natural Science Foundation of China (82170045 to J.H.); the Translational Medicine Cross Research Fund of Shanghai Jiao Tong University (ZH2018QNB29 to J.H.). This study was supported by the Macao Polytechnic University Internal Research Grant (RP/FCSD-02/2022 & RP/FCSD-03/2024) for H.T.

Data availability

DECEPTICON is compiled as an open-source R package available at https://github.com/Hao-Zou-lab/DECEPTICON.

Competing interests

The authors declare that they have no competing interests.

References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;**144**:646–74. https://doi.org/10.1016/j.cell.2011.02.013.
2. Xiao Z, Dai Z, Locasale JW. Metabolic landscape of the tumor microenvironment at single cell resolution. *Nat Commun* 2019;**10**:3763. https://doi.org/10.1038/s41467-019-11738-0.
3. Fridman WH, Pagès F, Sautès-Fridman C. *et al.* The immune contexture in human tumours: impact on clinical outcome. *Nat Rev Cancer* 2012;**12**:298–306. https://doi.org/10.1038/nrc3245.
4. Gajewski TF, Schreiber H, Fu YX. Innate and adaptive immune cells in the tumor microenvironment. *Nat Immunol* 2013;**14**:1014–22. https://doi.org/10.1038/ni.2703.
5. Lei X, Lei Y, Li JK. *et al.* Immune cells within the tumor microenvironment: biological functions and roles in cancer immunotherapy. *Cancer Lett* 2020;**470**:126–33. https://doi.org/10.1016/j.canlet.2019.11.009.
6. Liu X, Wu S, Yang Y. *et al.* The prognostic landscape of tumor-infiltrating immune cell and immunomodulators in lung cancer. *Biomed Pharmacother* 2017;**95**:55–61. https://doi.org/10.1016/j.biopha.2017.08.003.
7. Jochems C, Schlom J. Tumor-infiltrating immune cells and prognosis: the potential link between conventional cancer therapy and immunity. *Exp Biol Med (Maywood)* 2011;**236**:567–79. https://doi.org/10.1258/ebm.2011.011007.
8. Hadler-Olsen E, Wirsing AM. Tissue-infiltrating immune cells as prognostic markers in oral squamous cell carcinoma: a systematic review and meta-analysis. *Br J Cancer* 2019;**120**:714–27. https://doi.org/10.1038/s41416-019-0409-6.
9. Ni J, Si X, Wang H. *et al.* Prognostic biomarkers and immune cell infiltration characteristics in small cell lung cancer. *Cancer Pathog Ther* 2023;**1**:18–24. https://doi.org/10.1016/j.cpt.2022.09.004.
10. Smillie CS, Biton M, Ordovas-Montanes J. *et al.* Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* 2019;**178**:714–730.e22. https://doi.org/10.1016/j.cell.2019.06.029.
11. Pijuan-Sala B, Griffiths JA, Guibentif C. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 2019;**566**:490–5. https://doi.org/10.1038/s41586-019-0933-9.
12. Hashimoto K, Kouno T, Ikawa T. *et al.* Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. *Proc Natl Acad Sci USA* 2019;**116**:24242–51. https://doi.org/10.1073/pnas.1907883116.
13. Galon J, Costes A, Sanchez-Cabo F. *et al.* Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* 2006;**313**:1960–4. https://doi.org/10.1126/science.1129139.
14. Wu X, Ke X, Ni Y. *et al.* Tumor-infiltrating immune cells and PD-L1 as prognostic biomarkers in primary esophageal small

cell carcinoma. *J Immunol Res* 2020;**2020**:8884683. https://doi.org/10.1155/2020/8884683.
15. Petitprez F, Sun CM, Lacroix L. *et al.* Quantitative analyses of the tumor microenvironment composition and orientation in the era of precision medicine. *Front Oncol* 2018;**8**:390. https://doi.org/10.3389/fonc.2018.00390.
16. Lambrechts D, Wauters E, Boeckx B. *et al.* Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat Med* 2018;**24**:1277–89. https://doi.org/10.1038/s41591-018-0096-5.
17. Tsujikawa T, Kumar S, Borkar RN. *et al.* Quantitative multiplex immunohistochemistry reveals myeloid-inflamed tumor-immune complexity associated with poor prognosis. *Cell Rep* 2017;**19**:203–17. https://doi.org/10.1016/j.celrep.2017.03.037.
18. Becht E, Giraldo NA, Lacroix L. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol* 2016;**17**:218. https://doi.org/10.1186/s13059-016-1070-5.
19. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* 2017;**18**:220. https://doi.org/10.1186/s13059-017-1349-1.
20. Newman AM, Liu CL, Green MR. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015;**12**:453–7. https://doi.org/10.1038/nmeth.3337.
21. Racle J, de Jonge K, Baumgaertner P. *et al.* Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *elife* 2017;**6**:6. https://doi.org/10.7554/eLife.26476.
22. Hao J, Astle W, de Iorio M. *et al.* BATMAN—an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics* 2012;**28**:2088–90. https://doi.org/10.1093/bioinformatics/bts308.
23. Finotello F, Mayer C, Plattner C. *et al.* Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med* 2019;**11**:34. https://doi.org/10.1186/s13073-019-0638-6.
24. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* 2013;**29**:1083–5. https://doi.org/10.1093/bioinformatics/btt090.
25. Dong M, Thennavan A, Urrutia E. *et al.* SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief Bioinform* 2021;**22**:416–27. https://doi.org/10.1093/bib/bbz166.
26. Baron M, Veres A, Wolock SL. *et al.* A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst* 2016;**3**:346–360.e4. https://doi.org/10.1016/j.cels.2016.08.011.
27. Wang X, Park J, Susztak K. *et al.* Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun* 2019;**10**:380. https://doi.org/10.1038/s41467-018-08023-x.
28. Sturm G, Finotello F, Petitprez F. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* 2019;**35**:i436–45. https://doi.org/10.1093/bioinformatics/btz363.
29. Schelker M, Feau S, du J. *et al.* Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat Commun* 2017;**8**:2032. https://doi.org/10.1038/s41467-017-02289-3.
30. Avila Cobos F, Alquicira-Hernandez J, Powell JE. *et al.* Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat Commun* 2020;**11**:5650. https://doi.org/10.1038/s41467-020-19015-1.
31. Jin H, Liu Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biol* 2021;**22**:102. https://doi.org/10.1186/s13059-021-02290-6.
32. Decamps C, Arnaud A, Petitprez F. *et al.* DECONbench: a benchmarking platform dedicated to deconvolution methods for tumor heterogeneity quantification. *BMC Bioinformatics* 2021;**22**:473. https://doi.org/10.1186/s12859-021-04381-4.
33. Zou J, Deng F, Wang M. *et al.* scCODE: an R package for data-specific differentially expressed gene detection on single-cell RNA-sequencing data. *Brief Bioinform* 2022;**23**:bbac180. https://doi.org/10.1093/bib/bbac180.
34. Arnault G, Mony C, Vandenkoornhuyse P. Plant microbiota dysbiosis and the Anna Karenina principle. *Trends Plant Sci* 2023;**28**:18–30. https://doi.org/10.1016/j.tplants.2022.08.012.
35. Mahmoud N, Antson H, Choi J, Shimmi O, Roy K. Stress and Adaptation: Applying Anna Karenina Principle in Deep Learning for Image Classification. ArXiv, abs/2302.11380. 2023.
36. Zaneveld JR, McMinds R, Vega Thurber R. Stress and stability: applying the Anna Karenina principle to animal microbiomes. *Nat Microbiol* 2017;**2**:17121. https://doi.org/10.1038/nmicrobiol.2017.121.
37. Ma ZS. Testing the Anna Karenina principle in human microbiome-associated diseases. *iScience* 2020;**23**:101007. https://doi.org/10.1016/j.isci.2020.101007.
38. Galon J, Mlecnik B, Bindea G. *et al.* Towards the introduction of the 'Immunoscore' in the classification of malignant tumours. *J Pathol* 2014;**232**:199–209. https://doi.org/10.1002/path.4287.
39. Mlecnik B, van den Eynde M, Bindea G. *et al.* Comprehensive Intrametastatic immune quantification and major impact of Immunoscore on survival. *J Natl Cancer Inst* 2018;**110**:97–108. https://doi.org/10.1093/jnci/djx123.
40. Hao J, Liebeke M, Astle W. *et al.* Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nat Protoc* 2014;**9**:1416–27. https://doi.org/10.1038/nprot.2014.090.
41. Vallania F, Tam A, Lofgren S. *et al.* Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. *Nat Commun* 2018;**9**:4735. https://doi.org/10.1038/s41467-018-07242-6.
42. Fan J, Lyu Y, Zhang Q. *et al.* MuSiC2: cell-type deconvolution for multi-condition bulk RNA-seq data. *Brief Bioinform* 2022;**23**:bbac430. https://doi.org/10.1093/bib/bbac430.
43. Chu T, Wang Z, Pe'er D. *et al.* Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat Can* 2022;**3**:505–17. https://doi.org/10.1038/s43018-022-00356-3.
44. Hoek KL, Samir P, Howard LM. *et al.* A cell-based systems biology assessment of human blood to monitor immune responses after influenza vaccination. *PLoS One* 2015;**10**:e0118528. https://doi.org/10.1371/journal.pone.0118528.
45. Newman AM, Steen CB, Liu CL. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* 2019;**37**:773–82. https://doi.org/10.1038/s41587-019-0114-2.
46. Bocchino M, Zanotta S, Capitelli L. *et al.* Dendritic cells are the intriguing players in the puzzle of idiopathic pulmonary fibrosis pathogenesis. *Front Immunol* 2021;**12**:664109. https://doi.org/10.3389/fimmu.2021.664109.
47. Givi ME, Redegeld FA, Folkerts G. *et al.* Dendritic cells in pathogenesis of COPD. *Curr Pharm Des* 2012;**18**:2329–35. https://doi.org/10.2174/138161212800166068.

48. Cobos FA, Panah MJN, Epps J. *et al.* Effective methods for bulk RNA-seq deconvolution using scnRNA-seq transcriptomes. *Genome Biol* 2023;**24**:177. https://doi.org/10.1186/s13059-023-03016-6.
49. Pfister S, Kuettel V, Ferrero E. granulator: Rapid benchmarking of methods for in silico deconvolution of bulk RNA-seq data. R package version 1.16.0. 2025. https://bioconductor.org/packages/granulator. https://doi.org/10.18129/B9.bioc.granulator.
50. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021. https://www.R-project.org/.

Briefings in Bioinformatics, 2025, **26(3)**, bbaf234
https://doi.org/10.1093/bib/bbaf234
Problem Solving Protocol