Methods in
Molecular Biology 2959

Springer Protocols

Sweta Rani
Lukasz Skalniak *Editors*

IMMUNO-model in Cancer

Methods and Protocols

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Funded by
the European Union

OPEN ACCESS

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, UK

For further volumes:
http://www.springer.com/series/7651

For over 35 years, biological scientists have come to rely on the research protocols and methodologies in the critically acclaimed *Methods in Molecular Biology* series. The series was the first to introduce the step-by-step protocols approach that has become the standard in all biomedical protocol publishing. Each protocol is provided in readily-reproducible step-by-step fashion, opening with an introductory overview, a list of the materials and reagents needed to complete the experiment, and followed by a detailed procedure that is supported with a helpful notes section offering tips and tricks of the trade as well as troubleshooting advice. These hallmark features were introduced by series editor Dr. John Walker and constitute the key ingredient in each and every volume of the *Methods in Molecular Biology* series. Tested and trusted, comprehensive and reliable, all protocols from the series are indexed in PubMed.

IMMUNO-model in Cancer

Methods and Protocols

Edited by

Sweta Rani

Department of Science, South East Technological University, Waterford, Cork, Ireland

Lukasz Skalniak

Department of Organic Chemistry, Jagiellonian University, Kraków, Poland

Humana Press

Editors
Sweta Rani
Department of Science
South East Technological University
Waterford, Cork, Ireland

Lukasz Skalniak
Department of Organic Chemistry
Jagiellonian University
Kraków, Poland

European Cooperation in Science and Technology

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-0716-4733-2 ISBN 978-1-0716-4734-9 (eBook)
https://doi.org/10.1007/978-1-0716-4734-9

© The Editor(s) (if applicable) and The Author(s) 2026
Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.
The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Humana imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature.
The registered company address is: 1 New York Plaza, New York, NY 10004, U.S.A.

If disposing of this product, please recycle the paper.

Preface

Immuno-model is a model invented to study the immune system. This book describes the computational and experimental models that help researchers understand the responses of the immune system in cancer and test experimental immuno-oncology approaches.

Macrophages can adapt to different phenotypes in response to signals from the microenvironment. This book on immuno-model describes methods to profile polarization in macrophages using ELISA. ELISA is widely used in immunology to detect proteins, antibodies, antigens, or hormones in a sample, including immune checkpoint inhibitors. Immune checkpoints are regulatory molecules that control the activation and intensity of immune responses. Glycosylation is the addition of carbohydrate groups to the proteins. Glycosylation of the immune checkpoint proteins not only promotes immune evasion in tumor cells but also holds therapeutic implications. New immune checkpoint inhibitors are warranted for better cancer treatment. One of the chapters evaluates immune checkpoint inhibitors. One of the most studied immune checkpoints is PD-1/PD-L1. There is one chapter detailing the interaction of PD-1/PD-L1 and T cells. Immune checkpoint inhibitors have significantly improved survival rates in cancers but there are patients who do not respond to these treatments. Induction of immunogenic cell death is another therapeutic option for cancer patients.

A lot of research is underway to study tumor immune microenvironment. There are several well-established in vitro models to study interaction between immune cells and cancer cells and these in vitro models are still evolving. Cells can be co-cultured using cell culture inserts or can be grown as 3D spheroids. 3D co-culture model can be used to study the interaction of immune cells and cancer cells to mimic the in vitro microenvironment. Cells can be grown in 3D using different techniques, and one of the techniques is using scaffolds derived from cancer patients. One of the chapters explores immunocompetent preclinical mouse models to study primary and metastatic brain cancer. CAR T-cell therapy is still evolving, and one of the chapters describes the method to generate CAR T-cells.

Bioinformatics has vast applications and plays a central role in immunology by enabling the analysis of large-scale datasets. Deconvolution analysis can be used to study the epigenetic dysregulation in human tumors and the tumor ecosystem. Computational methods can also be used to study the mutated peptides called neoantigens. Bioinformatics allows us to identify therapeutic targets and develop precision immunotherapies.

Waterford, Cork, Ireland — *Sweta Rani*
Kraków, Poland — *Lukasz Skalniak*

Chapter 16

Deciphering the Tumor Microenvironment Composition Using Bulk Transcriptomics: A Guide to Recent Advances and Open Challenges

Sotiris Ouzounis, Donya Zojaji, Sandra García-Mulero, Marco Barreca, Paolo Gandellini, Theodora Katsila, Rebeca Sanz-Pamplona, and Maurizio Callari

Abstract

Tumors are complex ecosystems comprising diverse cell types actively participating to carcinogenesis, tumor progression, and treatment response. Understanding the tumor microenvironment (TME) dynamics has become of primary importance, especially with the increasing clinical implementation of immunotherapy. Low and high-throughput single cell and spatial technologies are providing high-resolution strategies for the study of the tumor ecosystem. However, their cost and complexity limit widespread use. Bulk transcriptomics has become a widely used strategy to obtain the expression profile of large cohorts of tumors or preclinical models. Several methods implementing a deconvolution analysis have been developed to estimate from bulk transcriptomics the prevalence of multiple cell types to reconstruct the tumor ecosystem composition.

In this chapter, we introduce deconvolution analysis, the main steps, the recent advancements, and open challenges. Our emphasis lies on robust benchmarking methodologies, highlighting the importance of clear parameter definition and appropriate metric selection for reliable results across different software tools.

Using CIBERSORTx and BayesPrism, we conduct a practical analysis on triple-negative breast cancer (TNBC) datasets from The Cancer Genome Atlas (TCGA) dataset. We illustrate the impact of various factors such as preprocessing methods, reference datasets, and software choice on deconvolution outcomes.

Integrating insights from benchmarking analyses and real-world applications, we provide guidance to optimize and control for the quality of deconvolution analysis, weighting both its potential and limitations. Deconvolution analysis can contribute to unravelling the complexities of the tumor microenvironment, but further research is needed to enhance accuracy and reproducibility.

Key words Deconvolution, Cancer, Microenvironment, Bulk, Transcriptomics, TME, Immune, Cell type, Challenges

---

Sotiris Ouzounis, Donya Zojaji and Sandra García-Mulero are Co-First authors.

Rebeca Sanz-Pamplona and Maurizio Callari are Co-Last authors.

Sweta Rani and Lukasz Skalniak (eds.), *IMMUNO-model in Cancer: Methods and Protocols*, Methods in Molecular Biology, vol. 2959, https://doi.org/10.1007/978-1-0716-4734-9_16, © The Author(s) 2026

1 Introduction

Tumors are highly heterogeneous entities including cancer cells, but also non-tumoral cells embedded in the tumor microenvironment (TME) [1], a complex network of multiple molecules and cell types [2].

A major role has emerged for the immune system infiltrating neoplastic lesions. Active research is being carried on the role of immune cells in carcinogenesis, cancer progression and response to therapy. The immune system plays a key role in cancer prevention and elimination, in a process known as immune surveillance. For example, through the recognition of neoantigens, cancer cells could be attacked and eliminated by effector T cells [3]. However, the immune system has a dual role in cancer. Tumors can also recruit immune cells that provide an immunosuppressive tumor microenvironment. Moreover, the stromal cells resident in the tissue can also have effects on tumor growth, like cancer associated fibroblasts and endothelial cells. Understanding the cross talk between cancer, stromal and immune cells is a hot spot in cancer research [4]. Therefore, to study cancer, it is crucial to identify the immune and stromal composition of a tumor.

A number of old and new techniques allows the study of the TME. New sequencing approaches such as single cell RNA-seq and spatial transcriptomics have been developed allowing direct high-resolution measurements of distinct cell-type prevalence and phenotype [5]. However, these techniques are quite demanding in terms of costs, computational power, and expertise. In addition, sample preparation is quite cumbersome and time-consuming.

Deconvolution methods have emerged as indirect techniques to quantify immune and other TME cell infiltration. While these methods are not able to achieve the resolution of single-cell techniques, they can be applied to widely available bulk transcriptomics data estimating, for example, the TME composition in large cancer sample cohorts. This process of cell type quantification is widely used in immunogenomics, and many statistical methods have been developed for quantification of immune and stromal cell types from bulk expression data obtained by RNA-seq and microarrays [6] (Fig. 1).

Scientific literature in cancer research has plenty of examples demonstrating the utility of deconvolution methods applied to transcriptomics data. Applications include generation of new hypotheses and searching for new biomarkers, among others. For example, in a meta-analysis by Kamal et al., four independent transcriptomics datasets were interrogated using deconvolution methods. About 22 immune cell types were inferred, and the association between immune infiltration by each cell type and relapse-free survival was assessed in colorectal cancer. As a result,

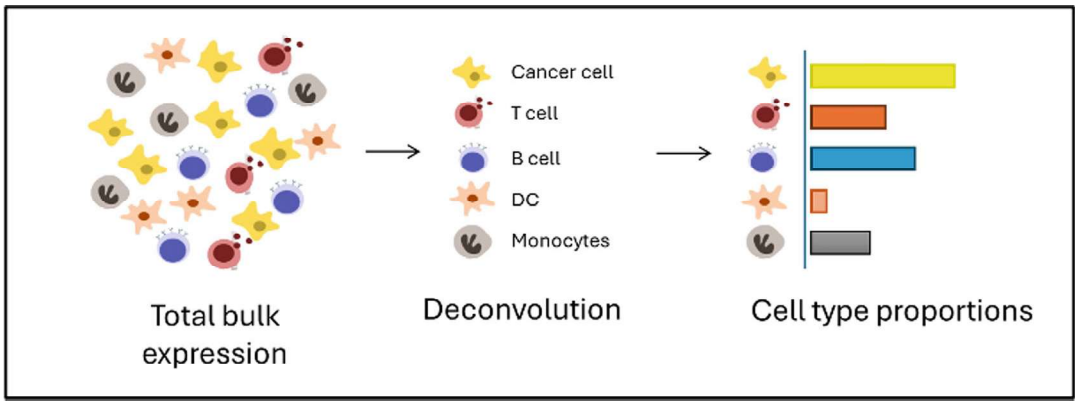

Fig. 1 Concept of immune cell infiltration estimation. From the total bulk RNA expression, it is possible to quantify the prevalence of distinct cell types infiltrating the tumor microenvironment

immune cell-type infiltration was found to be a better predictor of disease relapse than other expression-based biomarkers [7]. Specifically, CD4 and CD8 T cells and NK cells were found to be associated with better prognosis in this kind of tumor. Indeed, tumor infiltrating lymphocytes (TILs) have been associated with good prognosis in several cancer types [8]. On the contrary, infiltration of regulatory T cells (Tregs) has been described as a bad prognosis biomarker [9]. Immune cell infiltration could be also useful to generate new hypotheses. In a work by García-Mulero et al., metastatic samples from different cohorts were scored based on immune cell infiltration. Then, a cluster analysis was done, and three groups of samples emerged that were associated with immunotherapy response [10].

2 Available Methods and Key Steps of the Deconvolution Analysis

Systematic reviews summarising and benchmarking all the available tools have been reported elsewhere [11–13], but a summary of the most widely used open-source algorithms and their main features are listed in Table 1.

Deconvolution algorithms can be classified in various ways. Following the approach by Im and Kim [14], deconvolution algorithms can be categorized based on their methodology, prior knowledge of cell types, and methods of output. Based on the methodology used for inferring cell types, deconvolution tools can be divided into two categories: gene signature-based and fraction-based. Gene signature-based methods rely on the enrichment analysis of gene signatures. These methods are very useful for comparisons between phenotypes but cannot quantify inter-sample differences in cell-type abundance [15]. Fraction-based tools require a predefined reference matrix, which consists of expected

Table 1
Summary descriptive of methods for immune system infiltration estimation using transcriptomic profiles.

Tool	Approach	Cell fractions	Statistical method	Reference required	Cell-types	Applications	Reference
CIBERSORTx	Fraction-based	Relative/Absolute	Support vector regression	Bulk and scRNA-seq	Custom	Immune profiling of tumors	[29]
EPIC	Gene signature	Relative/Absolute	Constrained least squares regression	Bulk RNA-seq	Custom and predefined (six immune, two stromal, and unknown)	TME characterization	[36]
MCPcounter	Gene signature	Relative	Mean of marker gene expression	Bulk RNA-seq	Predefined (8 immune and 2 stromal)	TME characterization	[37]
TIMER	Fraction-based	Relative	ssGSEA	Bulk RNA-seq	Predefined (6 immune cells-types)	Immune profiling of tumors	[38]
quanTIseq	Fraction-based	Absolute	Constrained least squares regression	Bulk RNA-seq	Predefined (10 immune cell types)	Immune profiling of tumors	[39]
BayesPrism	Fraction-based	Relative	Bayesian framework	Bulk RNA-seq/scRNA-seq	Custom	High resolution for high granularity	[30]
ConsensusTME	Gene signature	Relative	ssGSEA	Bulk RNA-seq	Custom and predefined from previous methods (18 cell types)	Combine multiple methods for comprehensive TME characterization	[40]

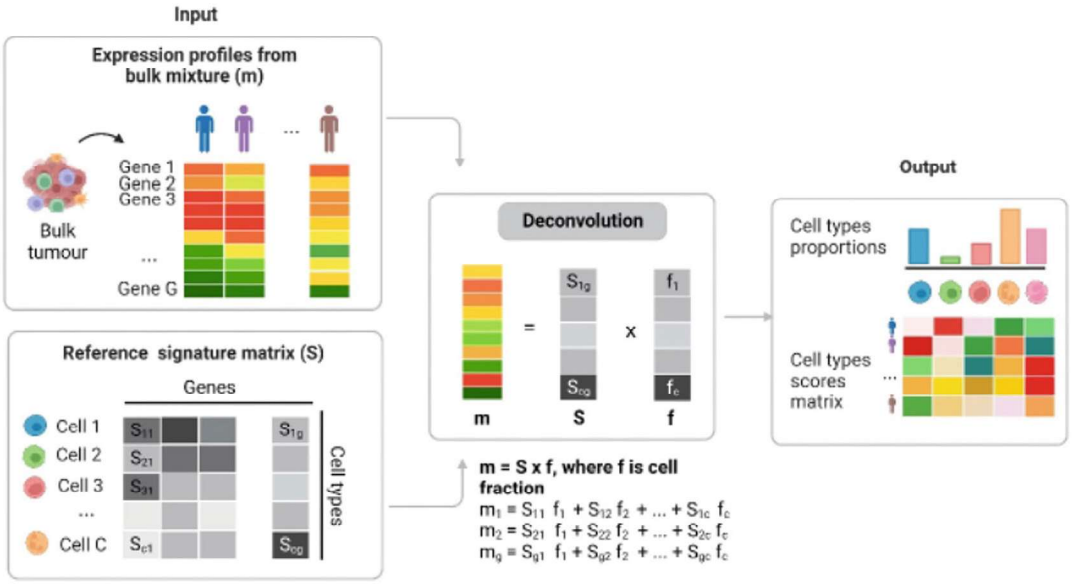

Fig. 2 Overview of deconvolution algorithms. Deconvolution from bulk transcriptomic data requires an expression matrix (m) from the bulk tissue and the cell-type specific reference signature matrix (S). The contribution of the different cell types can be inferred by regression models. The output of the deconvolution is a matrix of proportion scores per cell type and sample

values of gene expression for each cell type. This matrix is used to dissect the contribution of each signature profile to the aggregated bulk level of signal [16]. Furthermore, supervised or semi-supervised deconvolution methods can also be grouped based on the type of prior knowledge required. These include methods that rely on marker gene expression profiles, those that leverage single-cell RNA sequencing data to derive cell-type signatures, and those that account for gene expression variability within cell types.

Quantification methods can also be divided into two groups based on the abundance level: relative fraction or absolute fraction (Table 1). Relative methods estimate the relative abundance of each cell-type relative to each other, but do not provide information about the absolute quantification. The output of relative methods usually consists of enrichment scores and are useful for comparison across different samples or conditions (e.g., between different tumor types or disease states). Absolute scores, which account for the total proportions within a sample, provide more accurate and precise estimates of cell abundances, and allow both intra- and intercomparison of cell types [6].

Workflow in Fig. 2 explains fraction-based approaches that use an a priori defined reference matrix of expected values (S) and a gene expression matrix (m) from the interrogated sample. The bulk tissue representation (m) would result as the product of multiplying the reference signature (S) by the proportion contribution of each

cell type (f). To dissect the contribution of each cell type to the bulk signal, different statistical approaches can be performed, from linear regression models to more sophisticated deep learning algorithms [16].

Below, we review the three main steps for performing comprehensive deconvolution pipelines.

2.1 Input Data Preparation: Normalization and Data Transformation

Optimal data preprocessing and normalization are key aspects for the correct performance of deconvolution analysis. In general, log transformation is widely used in RNA-seq data analysis. However, deconvolution algorithms can be affected by data transformations [17]; Cobos et al. performed a benchmark where they evaluated different transformation methods (linear, Log, Sqrt and VST) and found that all deconvolution methods performed at best when applied to linear data. Preprocessing steps like normalization are potentially crucial when handling gene expression data. For this reason, they tested 20 different normalization methods, and found that the choice of normalization strategy has a minor impact in certain deconvolution algorithms. Overall, linear TPM (transcripts per million) is a suitable RNA-seq input data in many deconvolution methods. However, some deconvolution methods expect the data to be normalized (or not) and with a specific strategy, thus it is recommended to carefully read the instructions and recommendations given by the authors [18].

2.2 Selection of the Most Appropriate Method and Reference

The choice of methodology depends on the research question and the goals of the analysis. It is a crucial step in the deconvolution process since different methods could show high disparities in their performance and in the obtained results (see section. 4). Many factors can contribute to this variability, such as the statistical algorithm and the quality of the reference signatures, as discussed below [18].

Often, the proposed algorithms come with one or multiple reference matrices or signatures, which can differ considerably from each other and represent the main source of variability. These reference signatures can vary in the gene markers selected, the number of cell types included (from six to dozens), the diversity of cells included (only immune cells or accounting also for stromal components and/or malignant cells), and the level of granularity (i.e., the specificity of cell-types; e.g., the different subtypes of T cells) [16]. Reference signatures can be generated from bulk RNA-seq (usually derived by flow cytometry sorted populations or in vitro cultured cells) or from single-cell RNA-seq. Moreover, cell-type-specific expression profiles could be generated starting from different tissues (PBMCs, healthy organs, tumor tissues, etc.), as well as different model organisms.

The choice of the most appropriate reference matrix or signatures has a major impact on the quality of the deconvolution results,

as shown in the following sections. Ideally, the reference cell types should represent all cell types present in the bulk sample and their expression profile should be generated from tissues as similar as possible to the interrogated bulk sample. Deconvolution methods can overestimate (spillover effect) or even give positive results for absent cell types in the samples (background fraction prediction). This can happen when working with tumors with low immune infiltration and can lead to wrong conclusions. To deal with this problem, some methods like EPIC have added an "Unknown" category [13].

It is important to have the opportunity to use a custom reference matrix or signatures, possibly derived from the increasingly available single-cell RNA-seq data matching the bulk transcriptomics tissue. Moreover, it is common in the context of cancer research to use model organisms and derive nonhuman molecular data. The possibility of using custom references can consequently enable the opportunity to apply data deconvolution in this setting.

2.3 Output and Expected Results

The output of deconvolution algorithms consists of a matrix of proportion scores (Fig. 2), which can be further used for statistical comparisons among phenotypic groups, association studies and clustering analysis, among others. It is important to consider that the output can differ between different deconvolution methods due to the lack of standardized formats, normalization procedures, and scaling factors. Therefore, caution must be paid when attempting to integrate or compare outputs from different deconvolution methods, and careful consideration of methodological differences and validation against independent datasets may be necessary for robust interpretation. Often, additional metrics (e.g., root mean square error or RMSE and p values) providing some indication on the quality of the deconvolution process are provided.

3 Challenges in the Development and Benchmarking of Deconvolution Methods

Several challenges arise during the analysis of bulk transcriptomics from tumor samples using deconvolution methods. Over recent years, various approaches have been proposed and multiple tools have been developed to address these challenges [11, 19, 20]. Many tools aim to overcome known limitations of deconvolution, each offering unique features and benefits. However, despite the development of more robust and efficient tools, a major challenge has yet to be addressed. The challenge is to establish agreement among different tools, which is essentially affected by multiple different parameters of the deconvolution approaches. The agreement among computational tools can be generally defined as the consistency of results obtained from different software when applied to the same data. Therefore, the whole challenge

is to design and perform valid benchmarking among existing methods. To achieve this, the parameters of the benchmarking must be clearly defined, and appropriate metrics should be established to measure the agreement among tools and to identify parameters that may introduce bias into the consistency of results. Benchmarking is also essential during the development of deconvolution tools and the strategy used could impact the final performance and comparability with other methods.

Benchmarking process typically involves evaluating the accuracy of tools and various other parameters, such as computational resources required, processing time, scalability, and ease of use [11]. Common methods for evaluating deconvolution algorithms make use of (1) simulated bulk (or pseudo-bulk) data from single-cell RNA-seq, (2) bulk expression profiles from both pure and mixed cell lines, (3) data from the same tissue samples analyzed with both bulk and single-cell RNA-seq, (4) bulk transcriptomics and flow-cytometry data, and (5) bulk transcriptomics paired with clinical data. In all five cases above, the input data are bulk transcriptomics data, yet the ground truth data used for the evaluation of each method differs [11].

In the first case, single-cell data is used as ground truth and a dataset of bulk samples is simulated according to the cell type composition of the ground truth. Deconvolution accuracy is then established by quantitatively comparing the cell-type proportions of each pseudo-bulk sample with the ground truth. This approach is rather biased since simulated data are produced based on several assumptions that are study specific. Currently, several computational tools exist for simulating scRNA-seq count matrices [21–23], enabling the creation of "gold-standard" datasets. Nevertheless, it is important to recognize that artificially generated scRNA-seq data for constructing bulk mixtures might not completely capture the complexity of real biological data, potentially resulting in biased and overestimated performance evaluation of deconvolution algorithms [12]. When an algorithm is designed using real data, it usually exhibits better accuracy, as in the second case where different cell lines are mixed in predefined rations, to generate bulk samples. In such instances, the deconvolution accuracy is measured by comparing the estimated cell-type proportions in bulk samples with the expression profiles of pure cell lines. Even though these methods offer a more realistic approach, in vitro datasets are low throughput (small sample size, few cell types), which can make them prone to lack of generalization. Consequently, this may limit agreement with other tools. Similar limitations arise when algorithms are designed based on both bulk and single-cell RNA--seq data produced from the same tissue. In such cases, single-cell data serve as the ground truth for cell proportions to evaluate the algorithm's accuracy. Simultaneously, a subset of the single-cell data is used to construct the signature matrix. Thorough data handling

is required in this scenario since using the same single-cell data as both the ground truth and input for the signature matrix makes this approach susceptible to data leakage and, consequently, to overfitting. Specifically, while bulk transcriptomics data are typically gathered from intact tissue, the cell-type proportions are often evaluated using cell suspensions and methods like single-cell RNA sequencing or fluorescence-activated cell sorting. However, these methods can alter cell proportions, not accurately reflecting the original tissue composition. Consequently, comparing inferred cell-type proportions from bulk data with those measured from single-cell assays can lead to misinterpretations during the evaluation of deconvolution methods. To address this issue one solution would be to create consistent ground truth data for transcriptomics by dissociating the tissue specimen into cell suspensions. Then, use a portion of the suspension for bulk RNA sequencing and another portion for single-cell-based assays. This approach ensures that both bulk and single-cell data are obtained from the same starting material. Additionally, other technical factors can influence the creation of gold standard dataset such as the variations in cryopreserved samples, stored under different conditions, may yield different proportions of "live" cells compared to fresh samples [24]. In the fourth approach, flow cytometry is used as the ground truth to determine the number of each cell type in bulk samples. While this provides a reliable reference for constructing the signature matrix in reference-based algorithms, it is important to note that flow cytometry is typically used for blood samples. Consequently, methods validated solely with flow cytometry may overfit to blood samples and perform poorly when applied to other tissue types. In the last case, there is no ground truth data available—only clinical information such as survival time, disease status, and treatment response. Hence, only an indirect association between the estimated cell-type proportions from bulk data deconvolution and those reported either in literature or previous cases can be inferred. This approach may not be considered as a valid method for evaluating an algorithm since it lacks direct quantitative validation. However, it can be employed when other quantitative data are lacking. Considering the above, it is important to note that the method used to design and evaluate a tool can significantly impact its reproducibility and generalization ability.

To define a reference dataset either from in silico or in vitro data, several quality control steps should be followed to ensure the quality of the (raw) experimental data. Especially for sequencing data, either bulk or single cell, a common quality control step is the removal of genes full of zeroes or with zero variance read counts across all samples within the dataset. Additionally, for single cell sequencing, to assure high quality of raw data, cells with low quality of sequences, low numbers of unique molecular identifiers (UMIs),

and high fractions of ribosomal or mitochondrial content should be removed [25].

The generation of "gold-standard" datasets entails several challenges to be addressed, yet it is a crucial step for the quality control of the results. Toward the standardization of gold standard datasets and the benchmarking of the deconvolution algorithms, a recent study provides three "gold standard datasets" that were produced from imaging data with single-cell resolution [26]. This pipeline provides a high standard framework for benchmarking existing or new deconvolution algorithms while providing great reproducibility.

As mentioned above, a common factor that can influence the concordance of various tools analysing the same dataset is the preprocessing method employed. Sometimes, different tools require different preprocessing or normalization of the data and this could contribute to achieving discrepant estimations of cell type prevalences. Another parameter, often underestimated when comparing different algorithms, is the theoretical background according to which they are designed and the type of deconvolution algorithms used.

The evaluation of the agreement among different tools is also directly related to the metrics employed to quantitate the performance of the deconvolution method. The most common measure used to assess the accuracy of deconvolution algorithms is the correlation coefficient, which compares the computationally estimated cell-type proportions with the known proportions from the ground truth data. Pearson correlation is typically used, while Spearman correlation can also be employed. Pearson correlation indicates a linear relationship between estimated and ground truth data, while the Spearman coefficient reflects a monotonic relationship. Moreover, another proposed metric is correlation deviation [27]. This measure requires the sample size and the computation of the Pearson correlation between the estimated and the known counts per cell type. The formula of correlation deviation is provided below:

$$\text{Correlation deviation} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(1 - r_i)^2}$$

where n denotes the amount of different immune cell types found in the samples, r_i denotes the Pearson correlation coefficient for each immune cell type i. Additional robust metrics used for the evaluation of deconvolution methods RMSE and mean absolute error (MAE). These are two well-known measures for evaluating the accuracy of a computational tool, yet their main difference from a mathematical perspective is sensitivity to outliers. Outliers in the estimated values can greatly impact RMSE, while MAE is less sensitive to such data. However, the RMSE metric seems to lack

robustness when the concentrations of the cell types have a large variance in the bulk samples [28]. MAE is a robust measure in various settings; however, it only informs us about the absolute error of a method. Therefore, if used alone, it can be misleading for the performance of a deconvolution method applied to cell types with similar cell proportion [11]. It is evident that the evaluation of deconvolution methods relies on various metrics, each with distinct implications for accuracy and robustness, and the choice of metric can significantly influence the assessment of method performance.

4 Field Test of Deconvolution Methods

In this section, we aim at providing a practical example on how to perform a deconvolution analysis and examples of the impact of the software, processing and reference matrix on the final results. We first describe all steps leading to CIBERSORTx [29] analysis and then compare the results with those obtained using BayesPrism [30]. In a recent meta-analysis by Garmire et al., where they review on different benchmarking efforts, CIBERSORTx was one of the methods recommended by independent studies [12]. This method is widely used by the community thanks to its user-friendly web tool, which allows researchers to run it without having a computational background. Tran et al. evaluated the performance of nine different deconvolution methods and found that BayesPrism shows the best overall performance and prediction accuracy for nine major cell types in breast cancer samples, including normal epithelial, cancer epithelial, T cells, B cells, myeloid, endothelial, cancer-associated fibroblasts (CAFs), perivascular-like (PVL), and plasmablasts [13].

4.1 Example Analysis Using CIBERSORTx in TCGA

Here, we describe the main steps leading from pre-processed transcriptomic data to the deconvolution output. We will apply CIBERSORTx [29] in the context of TNBC, using publicly available data from The Cancer Genome Atlas (TCGA) [31] as bulk transcriptomics dataset to deconvolute and a single-cell RNA sequencing (scRNAseq) dataset [32] to derive the reference signature matrix.

Wu dataset [32] contains a total of 26 breast cancer patients belonging to all subtypes, for a total of over 100,000 cells, which have been assigned to 9 major cell types. The raw scRNA-seq of this dataset is available in the European Genome-Phenome Archive (EGA) under the accession code EGAS00001005173. The processed scRNA-seq data of this dataset is deposited in the Gene Expression Omnibus (GEO) with the accession number GSE176078. We manually downloaded the TAR format of processed transcription profiles from the supplementary file table (GSE176078_Wu_etal_2021_BRCA_scRNASeq.tar.gz). The

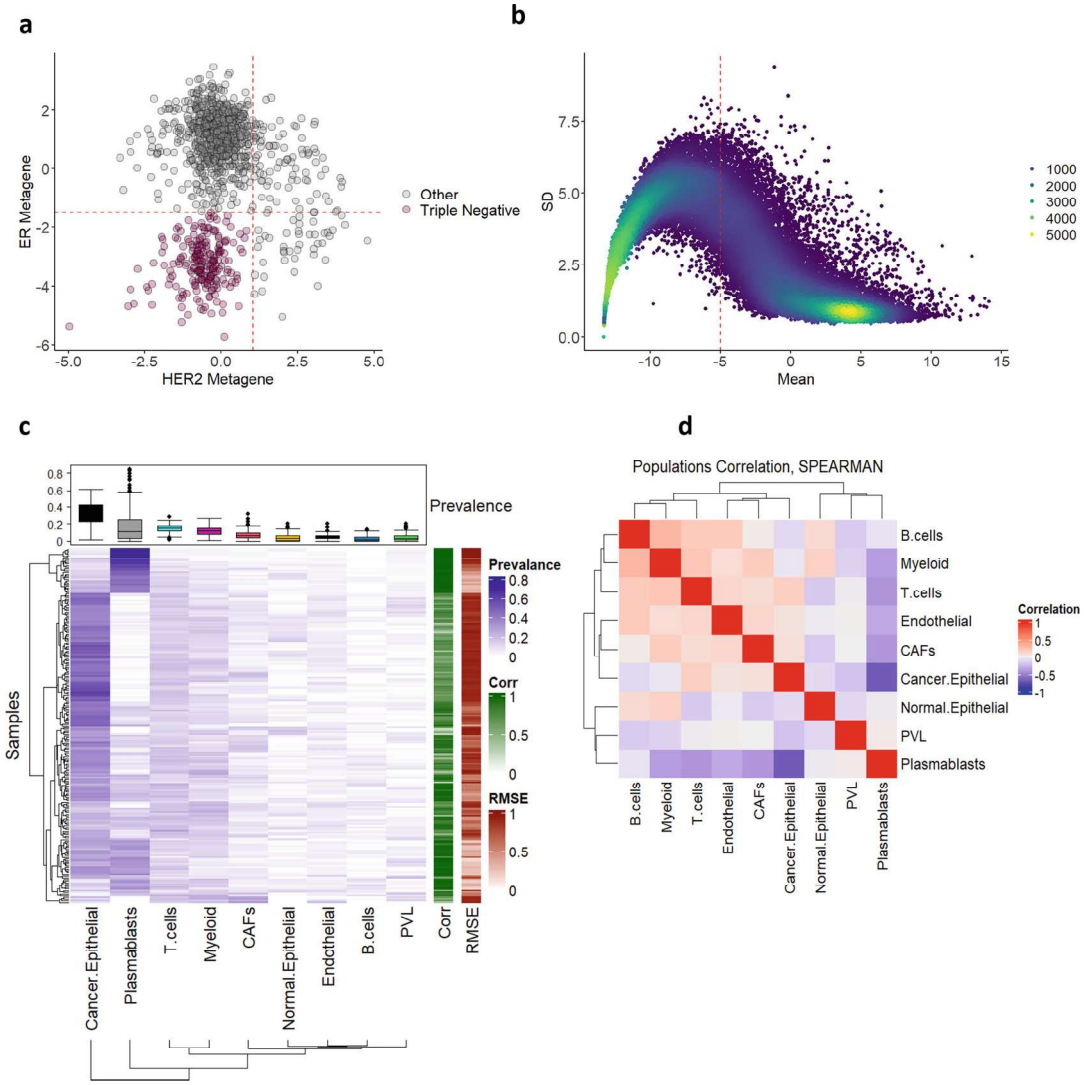

Fig. 3 TNBC TCGA data preparation and CIBERSORTx deconvolution**.** (**a**) Scatterplot of ER and HER2 metagenes used to identify the ER-HER2- or TNBC subtype in the breast cancer TCGA cohort. (**b**) Distribution of mean and standard deviation (SD) for each transcript in the TNBC TCGA cohort. Transcripts with a mean log2(TPM) value < -5 were discarded in downstream analyses. (**c**) Heatmap summarizing the output of CIBERSORTx deconvolution. Relative proportion for each cell type in each sample are shown. (**d**) Reciprocal correlation among all cell types detected in the TNBC TCGA dataset

patients' data was available in the supplementary file (1793222_Sup_Tab_1-11, Supplementary_table_1) of the article [32] or on the GEO page downloading it thought getGEO function (package GEOquery version 2.60.0).

The dataset was loaded and assembled in a Seurat object (version 5.0) in R/Bioconductor using the following code:

```
library(Seurat)

setwd("…/GSE176078_Wu_etal_2021_BRCA_scRNASeq")
Wu_em <- ReadMtx(
mtx = "matrix.mtx.gz", features = "features.tsv.gz",
cells = "barcodes.tsv.gz", feature.column = 1)

Wu_seurat <- CreateSeuratObject(counts = Wu_em)
```

Next, we added the patient and annotation data and selected TNBC tumors ($n = 9$) using the proper functions to manage Seurat object available on satijalab.org web site (https://satijalab.org/seurat/articles/essential_commands.html). We performed the standard workflow available on the same website to normalize, scale, and run the PCA and UMAP on the cell transcriptional profiles.

The TCGA dataset is accessible through the GDC Portal (https://portal.gdc.cancer.gov/) and/or by the R/Bioconductor TCGABiolinks package [33]. We downloaded TPM transcriptomic data related to breast cancer patients using the following commands:

```
library(TCGAbiolinks)

query <- GDCquery (
     project = "TCGA-BRCA",
     data.category = "Transcriptome Profiling",
     data.type = "Gene Expression Quantification",
     workflow.type = "STAR - Counts)

GDCdownload(query = query)

transcriptome <- GDCprepare(query = query)
TPM <- assays(transcriptome)$tpm_unstrand
```

After installing the TCGABiolinks package, we can access the data type and cancer type of interest by specifying them in the GDCquery() function. The data can later be downloaded from the GDC portal by using GDCdownload() function and stored and accessed locally by using GDCprepare(). Further details could be found at https://doi.org/10.18129/B9.bioc.TCGAbiolinks.

Sample filtering was required because the TCGA breast cancer transcriptomics dataset contains not only the expression profile from primary tumors but also from metastatic lesions and normal tissues. We used the sample ID to select solid tumor samples. Each sample in the TCGA repository has a unique barcode, which is the primary identifier of the biospecimen data within the TCGA project. For more information refer to https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/.

In TCGA-BRCA project, the solid tumor samples are identified by -01A- in the fourth block of the sample ID. The breast cancer project in TCGA contains 1231 samples and over 60,660 transcripts quantified. In this analysis, we specifically focused on the subset of TNBCc, identified by the lack of estrogen receptor (ER) expression and lack of human epidermal growth factor receptor 2 (HER2) amplification. Since immunohistochemistry (IHC)-based classification was not available for all cases, we used ER and HER2 metagene expression to identify the TNBC samples, as previously reported [34]. Metagene values were obtained by averaging the log10 TPM values of the genes belonging to the metagene. Based on the distribution of the metagenes, optimal thresholds were defined, and 213 samples were identified as TNBC (Fig. 3a).

To generate the input bulk transcriptomic matrix, a gene filtering was introduced. This had the double aim of removing genes with low/no expression and have unique gene symbols in the dataset. For all transcripts, we calculated the mean and standard deviation to characterize the overall data distribution (Fig. 3b). Transcripts with mean log10(TPM) lower than −5 were filtered out. Some duplicated gene symbols were present after this filtering and the one with highest IQR was selected. This way, 25,700 unique genes remained in the TNBC TCGA bulk transcriptomics. Expression profiles were exported as unlogged TPM.

The CIBERSORTx analysis was run online, following the extensive documentation present for details on data format and structure. The primary output of CIBERSORTx deconvolution is a file containing the prevalence of each cell type defined in the reference matrix in each sample. Some additional information (i.e., RMSE and correlation) is provided and can help with quality control. Output can be represented as a heatmap, as shown in Fig. 3c for the TNBC TCGA dataset. Such representation provides a general overview of the results, relative abundance of the different cell types and a qualitative indication on the presence of groups of patients with similar patterns in terms of cell-type prevalence. Additionally, we performed a reciprocal correlation analysis among cell types, which could inform us on the tendency to specific TME populations to coexist or being mutually exclusive (Fig. 3d).

4.2 Comparison of the Outputs from Different Tools

To compare the results of the deconvolution outputs using two different software, we analyzed the TNBC TCGA dataset with BayesPrism [30]. This software has been optimized to work with raw read count data instead of TPM, although the authors suggest the possibility to use the normalized data if the only available. To extract the raw read counts in the object downloaded using TCGA-Biolinks, we can use the function assays() and choose the data type we need. For a fare comparison, the same samples and genes included in the CIBERSORTx analysis were included here. We

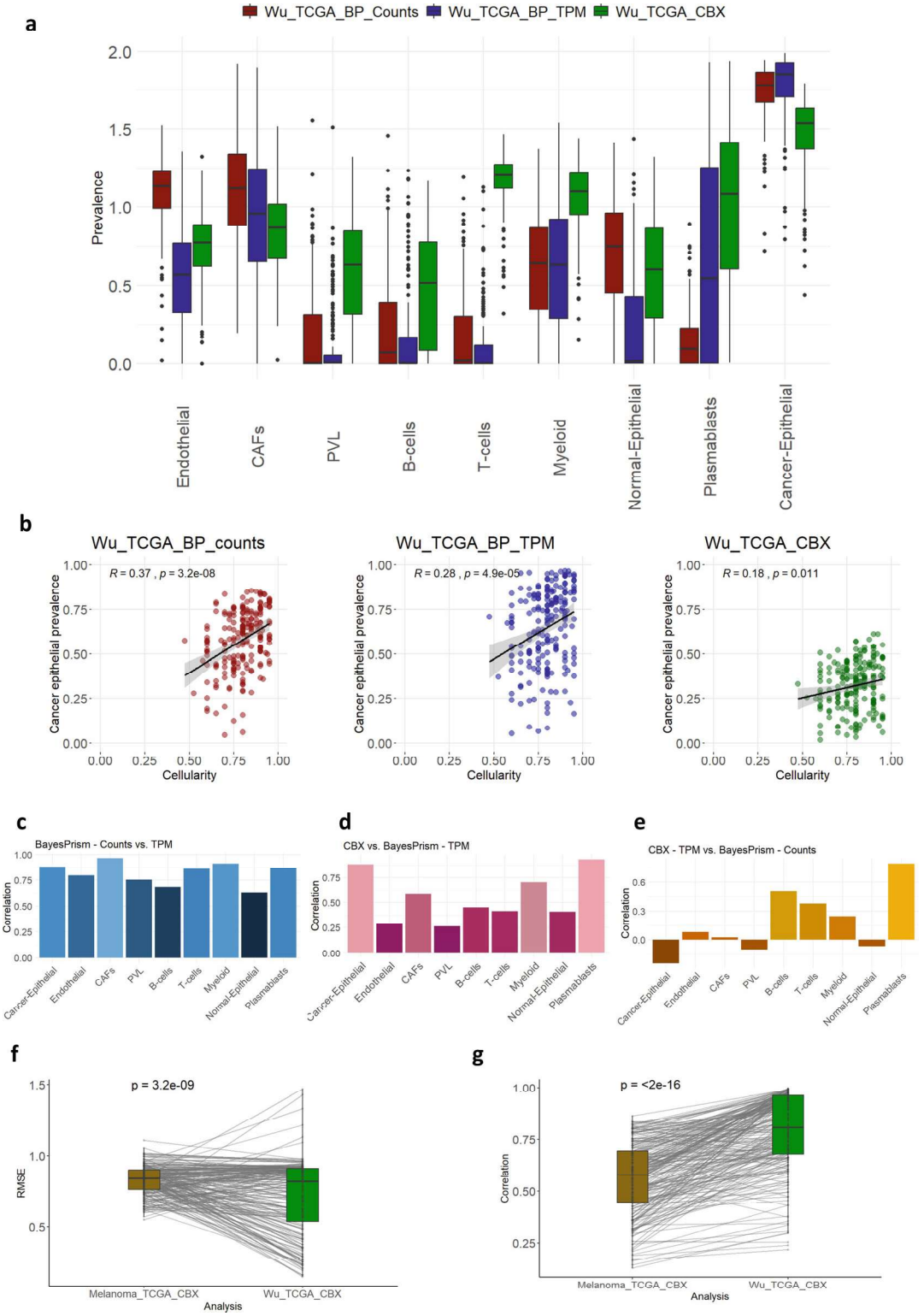

Fig. 4 Comparison of deconvolution results. (**a**) Boxplots showing cell-type prevalence distributions obtained with three deconvolution analyses applied to the tNBC TCGA dataset. (**b**) Correlation between pathologist estimated tumor cellularity and prevalence of cancer epithelial cells estimated by three deconvolution

remand to the online documentation for details on data format and structure and instructions on BayesPrism usage. For a thorough comparison with TPM-based deconvolution performed with CIBERSORTx, we additionally deconvoluted TNBC TCGA TPM data with BayesPrism.

As reported in Fig. 4a, prevalence of each cell type varied significantly depending on the software or data type used. For both methods cancer epithelial cells have the highest prevalence compared to other cell types; BayesPrism quantification of cancer cells were similar either starting from TPM or raw counts, but CIBERSORTx estimated lower prevalence. The same pattern can be seen for CAFs. A very noticeable difference can be seen in the T-cell prevalence estimation, where CIBERSORTx reported much higher prevalences than BayesPrism. For some cell types, namely endothelial, normal epithelial, and plasmablasts, prevalence estimated by BayesPrism starting from either raw counts or TPM was quite different. For the estimation of cancer epithelial cells, we could compare the deconvolution estimates with the tumor purity estimates obtained by the pathologists. Agreement between pathological and molecular estimation of tumor content has always shown low to moderate correlation [35]. Nevertheless, higher correlation values could be an indirect indication of higher performance of the deconvolution process. BayesPrism deconvolution starting from raw counts showed the highest agreement in our setting. This is in line with recent reports benchmarking multiple tools including BayesPrism [13] (Fig. 4b).

To further compare the results, we performed a correlation analysis between deconvolution results and for each cell type (Fig. 4c–e). Figure 4c, illustrates the correlation between BayesPrism results when using raw read counts or TPM as input. Overall, we found a good agreement with correlation reaching 0.9 for CAFs, but with the lowest correlation score being 0.7 for normal epithelial cells. Similarly, we compared CIBERSORTx results with BayesPrism results, for the latter starting either from TPM (Fig. 4d) or read counts (Fig. 4e). When comparing CIBERSORTx and BayesPrism outputs starting in both cases from TPM, cell-type correlations were remarkably lower for some cell types, reaching values around 0.25, except for plasmablasts, myeloids, and cancer epithelial (Fig. 4d). Correlation values went further down when CIBERSORT output was compared with BayesPrism output starting from read counts (Fig. 4e).

Fig. 4 (continued) analyses in the TNBC TCGA dataset. (**c**) Prevalence correlation for each cell type between BayesPrism results obtained starting from TPM or read counts. (**d**) Prevalence correlation for each cell type between CIBERSORTx and BayesPrism results using TPM as input. (**e**) Prevalence correlation for each cell type between CIBERSORTx results using TPM and BayesPrism results using read counts as input data

4.3 Effects of the Reference Data on the Output

Deconvolution tools usually provide reference matrices or signatures containing the expression profile of relevant genes driving the cell-type quantification in bulk samples. However, such reference data are often derived from quite different biological contexts. The context can affect cell phenotypes and, eventually, deconvolution performances when applied to bulk data from different biological contexts. To provide an example of how the choice of the reference matrix can affect the results, we repeated the CIBERSORTx analysis on TNBC TCGA data using the provided melanoma signature (https://doi.org/10.1126/science.aad0501). We compared the results in terms of RMSE and correlation values as provided by CIBERSORTx. Median RMSE was similar in the outputs obtained using the melanoma or the TNBC signature matrix, respectively. However, a strong reduction in RMSE was observed for 61.03% of the samples, while an increase was observed only in the 38.97% (Fig. 4f). Importantly, correlation between the original and the reconstructed bulk transcriptomic profile was significantly higher when using the TNBC matrix, matching the bulk transcriptome cancer type (median cor = 0.68 vs 0.84, $p < 2e\text{-}16$) (Fig. 4g).

5 Conclusions

Conducting a robust deconvolution analysis requires careful consideration of several factors. It is important to select the appropriate preprocessing method and normalization strategies, tailored to the specific dataset and software. Moreover, the choice of reference datasets and software tools significantly impacts deconvolution outcomes.

To ensure the quality of results, it is essential to validate findings using complementary experimental techniques and independent datasets whenever possible. Additionally, researchers should remain mindful of the limitations inherent in bulk transcriptomics-based deconvolution, including potential biases introduced by tissue heterogeneity, batch effects, and the reliance on predefined reference matrices. While deconvolution analysis offers valuable insights into tumor microenvironment dynamics, it is important to interpret results cautiously and consider the broader context of tumor biology.

Looking ahead, ongoing research efforts should focus on refining benchmarking methodologies, improving software tools, and addressing remaining limitations in deconvolution analysis. By advancing our understanding of the tumor microenvironment and enhancing the accuracy and reproducibility of deconvolution methods, we can pave the way for more effective precision oncology strategies and ultimately improve patient outcomes in cancer treatment.

Acknowledgments

This work has been supported by Fondazione Michelangelo (Milan, Italy) and Breast Cancer Research Foundation (grant BCRF 21-181), the Aragon Government (Group B29_23R), the Instituto de Salud Carlos III (ISCIII) grant PI22/01938, and by ASPANOA Foundation. This work is part of the CNS2022-136176 action, financed by MCIN/AEI/10.13039/501100011033 and for the European Union «Next Generation EU»/PRTR. This publication is based upon work from COST Action IMMUNO-model, CA21135, supported by COST (European Cooperation in Science and Technology). We would like to acknowledge, in memory of Sotiris, his early contributions to this work. His presence and spirit remain fondly remembered.

References

1. de Visser KE, Joyce JA (2023) The evolving tumor microenvironment: from cancer initiation to metastatic outgrowth. Cancer Cell 41: 374–403
2. Anderson NM, Simon MC (2020) The tumor microenvironment. Curr. Biol. 30:R921–R925
3. Lakshmi Narendra B, Eshvendar Reddy K, Shantikumar S, Ramakrishna S (2013) Immune system: a double-edged sword in cancer. Inflamm. Res. 62:823–834
4. Swanton C, Bernard E, Abbosh C, André F, Auwerx J, Balmain A et al (2024) Embracing cancer complexity: hallmarks of systemic disease. Cell 187:1589–1616
5. Dezem FS, Morosini NS, Arjumand W, DuBose H, Plummer J (2024) Spatially resolved single-cell omics: methods, challenges, and future perspectives. Annu Rev Biomed Data Sci 7(1)
6. Hackl H, Charoentong P, Finotello F, Trajanoski Z (2016) Computational genomics tools for dissecting tumour-immune cell interactions. Nat. Rev. Genet. 17:441–458
7. Kamal Y, Dwan D, Hoehn HJ, Sanz-Pamplona R, Alonso MH, Moreno V et al (2021) Tumor immune infiltration estimated from gene expression profiles predicts colorectal cancer relapse. Onco Targets Ther 10: 1862529
8. Brummel K, Eerkens AL, de Bruyn M, Nijman HW (2023) Tumour-infiltrating lymphocytes: from prognosis to treatment selection. Br. J. Cancer 128:451–458
9. Saleh R, Elkord E (2020) FoxP3+ T regulatory cells in cancer: Prognostic biomarkers and therapeutic targets. Cancer Lett. 490:174–185
10. García-Mulero S, Alonso MH, Pardo J, Santos C, Sanjuan X, Salazar R et al (2020) Lung metastases share common immune features regardless of primary tumor origin. J. Immunother. Cancer 8:e000491
11. Nguyen H, Nguyen H, Tran D, Draghici S, Nguyen T (2024) Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges. Nucleic Acids Res. 52:4761
12. Garmire LX, Li Y, Huang Q, Xu C, Teichmann SA, Kaminski N et al (2024) Challenges and perspectives in computational deconvolution of genomics data. Nat. Methods 21:391–400
13. Tran KA, Addala V, Johnston RL, Lovell D, Bradley A, Koufariotis LT et al (2023) Performance of tumour microenvironment deconvolution methods in breast cancer using single-cell simulated bulk mixtures. Nat. Commun. 14:5758
14. Im Y, Kim Y (2023) A comprehensive overview of RNA deconvolution methods and their application. Mol. Cells 46:99–105
15. Sturm G, Finotello F, List M (2020) Immunedeconv: an R package for unified access to computational methods for estimating immune cell fractions from bulk RNA-sequencing data. Methods Mol. Biol. 2120:223–232
16. Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH et al (2019) Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics 35:i436–i445
17. Zhong Y, Liu Z (2011) Gene expression deconvolution in linear space. Nat. Methods 9:8–9

18. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K (2020) Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat. Commun. 11: 5650
19. Finotello F, Trajanoski Z (2018) Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunol. Immunother. 67:1031–1040
20. Maden SK, Kwon SH, Huuki-Myers LA, Collado-Torres L, Hicks SC, Maynard KR (2023) Challenges and opportunities to computationally deconvolve heterogeneous tissue with varying cell sizes using single-cell RNA-sequencing datasets. Genome Biol. 24:288
21. Zappia L, Phipson B, Oshlack A (2017) Splatter: simulation of single-cell RNA sequencing data. Genome Biol. 18:174
22. Zhang X, Xu C, Yosef N (2019) Simulating multiple faceted variability in single cell RNA sequencing. Nat. Commun. 10:2611
23. Cao Y, Yang P, Yang JYH (2021) A benchmark study of simulation methods for single-cell RNA sequencing data. Nat. Commun. 12: 6911
24. Denisenko E, Guo BB, Jones M, Hou R, De Kock L, Lassmann T et al (2020) Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol. 21:130
25. Subramanian A, Alperovich M, Yang Y, Li B (2022) Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics. Genome Biol. 23:267
26. Sang-aram C, Browaeys R, Seurinck R, Saeys Y (2024) Spotless, a reproducible pipeline for benchmarking cell type deconvolution in spatial transcriptomics. Elife 12:RP88431
27. Miao YR, Zhang Q, Lei Q, Luo M, Xie GY, Wang H et al (2020) ImmuCellAI: a unique method for comprehensive T-cell subsets abundance prediction and its application in cancer immunotherapy. Adv Sci (Weinh) 7:1902880
28. Alonso-Moreda N, Berral-González A, De La Rosa E, González-Velasco O, Sánchez-Santos JM, De Las RJ (2023) Comparative analysis of cell mixtures deconvolution and gene signatures generated for blood, immune and cancer cells. Int. J. Mol. Sci. 24:10765
29. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F et al (2019) Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. 37:773–782
30. Chu T, Wang Z, Peer D, Danko CG (2022) Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. Nat Cancer 3:505–517
31. Hutter C, Zenklusen JC (2018) The cancer genome atlas: creating lasting value beyond its data. Cell 173:283–285
32. Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A et al (2021) A single-cell and spatially resolved atlas of human breast cancers. Nat. Genet. 53:1334–1347
33. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D et al (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 44: e71
34. Callari M, Cappelletti V, D'Aiuto F, Musella V, Lembo A, Petel F et al (2016) Subtype-specific metagene-based prediction of outcome after neoadjuvant and adjuvant treatment in breast cancer. Clin. Cancer Res. 22:337–345
35. Haider S, Tyekucheva S, Prandi D, Fox NS, Ahn J, Xu AW et al (2020) Systematic assessment of tumor purity and its clinical implications. JCO Precis. Oncol. 4:995–1005
36. Racle J, Gfeller D (2020) EPIC: a tool to estimate the proportions of different cell types from bulk gene expression data. Methods Mol. Biol. 2120:233–248
37. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F et al (2016) Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 17:218
38. Li B, Li T, Liu JS, Liu XS (2020) Computational deconvolution of tumor-infiltrating immune components with bulk tumor gene expression data. Methods Mol. Biol. 2120: 249–262
39. Plattner C, Finotello F, Rieder D (2020) Deconvoluting tumor-infiltrating immune cells from RNA-seq data using quanTIseq. Methods Enzymol. 636:261–285
40. Jimenez-Sanchez A, Cast O, Miller ML (2019) Comprehensive benchmarking and integration of tumor microenvironment cell estimation methods. Cancer Res. 79:6238–6246

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.