Comprehensive evaluation of cell-type quantification methods for immuno-oncology

- Gregor Sturm (1,2,*) [https://orcid.org/0000-0001-9584-7842], g.sturm@tum.de
- Francesca Finotello (3) [https://orcid.org/0000-0003-0712-4658], francesca.finotello@i-med.ac.at
- Florent Petitprez (4,5) [https://orcid.org/0000-0002-5117-6466], florent.petitprez@crc.jussieu.fr
- Jitao David Zhang (6) [http://orcid.org/0000-0002-3085-0909], jitao_david.zhang@roche.com
- Jan Baumbach (1) [https://orcid.org/0000-0002-0282-0462], jan.baumbach@wzw.tum.de
- Wolf H. Fridman (4), herve.fridman@crc.jussieu.fr
- Markus List (7) [https://orcid.org/0000-0002-0941-4168], markus.list@wzw.tum.de
- Tatsiana Aneichyk (2), aneichyk@pieris.com

(1) Chair of Experimental Bioinformatics, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Maximus-von-Imhof-Forum 3, 85354 Freising, Germany
(2) Pieris Pharmaceuticals GmbH, Lise-Meitner-Straße 30, 85354 Freising, Germany
(3) Biocenter, Division for Bioinformatics, Medical University of Innsbruck, 6020 Innsbruck, Austria
(4) Cordeliers Research Centre, UMRS_1138, INSERM, University Paris-Descartes, Sorbonne University, 75006 Paris, France
(5) Ligue Nationale contre le Cancer, Programme Cartes d'Identité des Tumeurs, 75013 Paris, France
(6) Roche Pharma Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland
(7) Big Data in BioMedicine, Chair of Experimental Bioinformatics, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Maximus-von-Imhof-Forum 3, 85354 Freising, Germany

* corresponding author

Abstract

The composition and density of immune cells in the tumor microenvironment profoundly influence tumor progression and success of anti-cancer therapies. Flow cytometry, immunohistochemistry staining, or single-cell sequencing is usually not available such that we rely on computational methods to estimate the immune-cell composition from bulk RNA-sequencing (RNA-seq) data. Various methods have been proposed recently, yet their capabilities and limitations have not been evaluated systematically. A general guideline leading the research community through cell type deconvolution is missing. We developed a systematic approach for benchmarking such computational methods and assessed the accuracy of tools for estimating a variety of immune cell types from bulk RNA-seq samples. We used a single-cell RNA-seq dataset of ~11,000 cells from the tumor microenvironment to simulate bulk samples of known cell type proportions, which allowed us to measure the accuracy of the computational methods. Moreover, we collected publicly available data that provide both RNA-seq and independent, gold-standard estimates of the cell type proportions. This allowed us to analyze and condense the results of a hundred of thousand predictions to provide an exhaustive evaluation across computational methods over ten cell types and ~1,400 samples from five simulated and real-world datasets. We demonstrate that computational deconvolution performs at high accuracy for well-defined high-quality signatures, demonstrating its practical utility of bulk RNA-seq to dissect immune-cell composition of tumor samples. In addition to guidelines for method selection (depending on the cell types of interest) we also offer an R package to allow the community to perform integrated deconvolution using different methods themselves (https://grst.github.io/immunedeconv). We suggest that future efforts should be dedicated to refining cell population definitions and finding reliable signatures.

Tumors are not only composed of malignant cells, but are embedded in a complex microenvironment within which dynamic interactions are built[1]. Notably, this tumor microenvironment (TME) comprises a vast variety of immune cells. Knowledge of immune cell content in cancer samples is invaluable for cancer-immunotherapy drug discovery as well as for clinical decisions about treatment options. The cellular composition of the immune infiltrate of tumors can shed light on the escape mechanisms that tumor cells use to evade the immune response. In clinical trials, it can be used to stratify patients for most suitable treatment options depending on the targeted cell type, hence increasing the overall chances of success, and ultimately accelerating access to improved treatment options[2].

Methods like fluorescence activated cell sorting (FACS) or immunohistochemistry (IHC)-staining have been used as gold standards to estimate the immune cell content within a sample[3]. However, each of the methods has its technical limitations and might not be applicable in certain situation. More specifically, FACS requires a large amount of material, hence limiting its application to tumor biopsies, and IHC provides an estimate from a single tumor slice, which may not be representative of a heterogeneous immune landscape of the tumor. Furthermore, both methods can utilize only a small number of cell type specific markers. More recently, single-cell RNA sequencing (scRNA-seq) is used to characterize cell types and states, yet for the time being it remains too expensive and laborious for routine clinical use. Moreover, cell type proportions might be biased in scRNA-seq data due to differences in single-cell dissociation efficiencies[4]. At the same time, methods for gene expression profiling, RNA-seq and microarrays, have been developed and optimized to be applicable in clinical settings, resulting in a plethora of transcriptomic datasets derived from patient tumor samples. More recently, technologies like Nanostring and *TruSeq RNA Access by Illumina* made expression profiling available even for formalin-fixed paraffin-embedded (FFPE) samples. However, these methods do provide transcriptomics data from heterogeneous samples considered as a whole, but no detailed information on their cellular composition, which therefore has to be inferred using computational techniques.

Computational methods for immune cell content estimation in tumor samples can, in general, be classified in two categories: marker gene-based approaches and deconvolution-based approaches (Table 1). Marker gene-based approaches utilize a list of genes that are characteristic for a cell type. These gene sets are usually derived from targeted transcriptomics studies characterizing each immune-cell type and/or from comprehensive literature search and experimental validation. By using the expression values of marker genes in heterogeneous samples, every cell type is quantified independently, either aggregating them into an abundance score (MCP-counter)[5] or by performing a statistical test

for enrichment of the marker genes (xCell)[6]. Deconvolution methods, on the other hand, formulate the problem as a system of equations that describe the gene expression of a sample as the weighted sum of the expression profiles of the admixed cell types (reviewed in [7]). By solving the inverse problem, cell type fractions can be inferred given a signature matrix and the bulk gene expression. In practice, this problem can be solved using □-Support Vector Regression (CIBERSORT)[8], constrained least square regression (quanTIseq and EPIC)[9,10] or linear least square regression (TIMER)[11].

While the available methods have been reviewed in terms of methodology[3,7,12,13], no quantitative comparison has been undertaken to demonstrate the power and limitations of the individual methods. Conducting a fair benchmark between immune-deconvolution methods is challenging, as the scores of the different methods have different properties and are not always directly comparable (Table 1). For instance, an inherent limitation of marker-based approaches is that they only allow to generate a semi-quantitative score, which enables a comparison between samples, but not between cell types[14]. Deconvolution allows, in principle, to generate absolute scores that can be interpreted as cell fractions and compared both inter- and intra-sample. In practice, CIBERSORT expresses results as a fraction relative to the immune-cell content and TIMER produces a score in arbitrary units that cannot be compared between cell types. Both quanTIseq and EPIC generate scores relative to the total amount of sequenced cells, being the only two methods generating an absolute score that can be interpreted as a cell fraction. CIBERSORT has recently been extended to the “absolute mode”, which provides a score that can be compared between both samples and cell types but still cannot be interpreted as a cell fraction[15]. From now on, we refer to this extension as “CIBERSORT abs”. quanTIseq comes as an entire pipeline that allows to directly analyze raw RNA-seq data (*i.e.* FASTQ files of sequencing reads), whereas the other methods only analyze pre-computed, normalized expression data. In this benchmark, we consider only its “deconvolution” module alone, acknowledging that the full quanTIseq pipeline can, in principle, provide higher performance. A limitation of TIMER and xCell is that the results of these methods depend on all samples that are analyzed in a single run, *i.e.* the same sample can have different scores, when submitted together with different other samples. In particular, xCell uses the variability among the samples for a linear transformation of the output score. TIMER uses COMBAT[16] to merge the input samples with reference profiles. This is particularly problematic with small datasets and hampers comparability and interpretability of the score. Moreover, xCell does not detect any signal when ran with few non-heterogeneous samples[17]. Nonetheless, all methods except CIBERSORT (relative mode) support between-sample comparisons and the performance of

the methods can be readily compared using Pearson correlation with the gold-standard measure of cell fractions.

Here we developed a systematic approach for benchmarking such computational methods and assessed the accuracy of estimating the abundance of seven immune cell types from bulk RNA-seq samples for seven state-of-the-art approaches that come together with signature matrices specifically designed for immuno-oncology, and provide immuno-oncologists good guidance on which method they can apply in their studies. We use a scRNA-seq dataset of more than 11,000 single cancer, stromal, and immune cells[18] to simulate bulk RNA-seq samples with known compositions and validate the benchmark results using three independent datasets that have been profiled with FACS[10,18,19]. Using single cell data not only circumvents the limited availability of RNA-seq samples with matched FACS or IHC profiles but also allows us to generate samples of arbitrary compositions, uniquely enabling us to assess minimal detection fractions and background predictions. We assess the performance of the methods by four metrics, namely: (i) predictive performance, (ii) minimal detection fraction, (iii) background predictions and (iv) spillover effects. Additionally, we take advantage of RNA-seq data from 10,273 primary tumor samples from The Cancer Genome Atlas (TCGA)[20] to assess the heterogeneity of cell-type signatures and relate the results to the observation of the spillover benchmark. Finally, we provide an R package, `immunedeconv`, integrating all of the methods for easy access. To facilitate easy benchmarking of newly developed methods and signatures, we provide a reproducible pipeline on GitHub, which can be easily extended to additional methods.

Results

Benchmark on simulated and true bulk RNA-seq samples reveals differences in method performance between cell types

We created 100 simulated bulk-tumor RNA-seq samples by aggregating 500 randomly-drawn, single immune and non-immune cells each following the strategy proposed by Schelker *et al.* for benchmarking CIBERSORT[18]. Additionally, we ensured that simulated bulk RNA-seq data is not subject to systematic biases (supplementary figures 1-4). As xCell, MCPcounter, and EPIC allows also to quantify cancer-associated fibroblasts (CAFs) and endothelial cells, we leveraged on the annotations available in this dataset to additionally benchmark the methods on non-immune cell types. We ran the methods on theses samples

and compared the estimated to the known fractions. The results are shown in figure 1a. All methods obtained a high correlation on B cells (r > 0.71), cancer associated fibroblasts (r > 0.79), endothelial cells (r > 0.94) and $CD8^+$ T cells (r > 0.76). Most methods obtain a correlation of r > 0.68 on macrophages/monocytes, NK cells and overall $CD4^+$ T cells. We observed poor performance in distinguishing regulatory- from non-regulatory $CD4^+$ T cells and in estimating dendritic cells (DCs).

As an additional validation, we assessed the performance of the methods on three independent datasets, which provide both bulk RNA-seq data and an estimate of immune cell contents using FACS (figure 1b-d and supplementary figure 5). In general, the results are highly consistent with the mixing benchmark, with the exception that the methods' performance on $CD8^+$ T cells is worse on the validation data considered altogether (figure 1c). This can be attributed to the fact that the variance of $CD8^+$ T cell contents between the samples is low and the correlations are therefore not meaningful (see figure 1b and supplementary figure 5). Moreover, in the simulation benchmark, only TIMER detects DC, while in Hoek's dataset[19], all methods except CIBERSORT detect a signal. This inconsistency, and the poor performance on DCs in general might be indicative of the still not well defined biological heterogeneity between DC subtypes.

Background predictions affect most methods

Next we investigated two related questions: at which abundance level do methods reliably identify the presence of immune cells (minimal detection fraction) and, conversely, what fractions of a certain cell type are predicted even when they are actually absent (background predictions). To this end, we again simulated bulk RNA-seq samples using the single cell dataset. For each immune cell-type, we spiked-in an increasing amount of cells of interest into a background of ~1,800 non-immune cells (fibroblasts, endothelial- and cancer cells) and measured a *background prediction* level as the fraction of predicted cells at zero immune infiltration, and the *minimal detection* level, at which the predicted immune cell content is significantly different from the background.

We observed that all deconvolution-based approaches almost always predict a minimal amount of immune cells even though they are absent (figure 2). This is expected for CIBERSORT (relative mode), as fractions are expressed relative to total immune cell content, but should not be the case for all other methods. The only exceptions are the low amount of background predictions of EPIC for B cells, Macrophages/Monocytes and NK cells and of quanTIseq for $CD8^+$ T cells. In contrast, xCell, which uses a statistical test for

enrichment of the maker-genes, is highly robust against background-predictions (score = 0 for all tested immune cell types) at a reasonable minimal detection fraction (0-1% infiltration depending on the immune cell type). Unfortunately, testing MCP-counter for its detection limit and background-predictions is not straightforward, as the method does not compute a score, but uses raw gene expression values. To make a statement about the absence of a cell type, a platform-specific null-model needs to be generated, which, in fact, is already provided by the authors for some microarray platforms, but not for RNA-seq. The authors also addressed the detection limit on RNA mixtures in their original publication[5]. In short, we observe that all deconvolution-based approaches are susceptible to background-predictions that might be due to the similarity of the signatures of closely-related cell types (*multicollinearity*) and/or to a lower cell-specificity of signature/marker genes. For $CD4^+$ T cells, which show a particularly high background prediction level, we identified five genes, *FHIT, GNG2, TSPYL1, DGKA* and *CD79B*, that are highly expressed in both $CD4^+$ T cells and the non-immune cells used as background. When excluding these genes from the signatures matrices, we could reduce the background prediction level of quanTIseq and EPIC by 54% and 36% respectively (figure 3b).

Spillover can be attributed to non-specific signature genes

Due to multicollinearity, it may happen that the abundance of a certain cell type leads to a method erroneously reporting higher abundance of another, here referred to as *spillover*. We assessed spillover effects using simulated bulk RNA-seq samples of pure immune cells and calculated an *overall noise ratio* for each method, *i.e.* the fraction of predictions that are attributed to a wrong cell type (figure 3a).

Spillover appears to primarily happen between $CD8^+$ and $CD4^+$ T cells and from DCs to monocytes/macrophages and B cells. To ensure that these effects are not driven by biases in the single-cell dataset, we validated our results on independent reference samples of pure immune cells (supplementary figure 6). We could confirm the spillover between $CD8^+$ and $CD4^+$ T cells and between DCs and macrophages/monocytes, but we could not detect substantial spillover between DCs and B cells by EPIC, quanTIseq and CIBERSORT. To exclude the possibility of erroneous annotation of single cells in the benchmark dataset, we demonstrate that the two cell-type clusters are both distinct and well-annotated (supplementary figure 7). We investigated which marker genes drive the spillover by checking which marker genes from the signature matrices of the three methods are over-represented in both the B cell and DC subpopulation of the single cell dataset. We identified six genes, *TCL1A*, *TCF4*, *CD37*, *SPIB*, *IRF8* and *BCL11A*, that are both expressed in the B

cell and DC subpopulations on the mRNA level. Indeed, in the LifeMap Discovery database[21], all six genes are annotated as being expressed in both B cells and plasmacytoid dendritic cells (pDCs). When excluding these genes from the deconvolution matrices, the spillover effect is substantially reduced by 55% for EPIC and 78% for quanTIseq (figure 3c). Given that the single cell dataset contains pDCs, quanTIseq has been trained on myeloid dendritic cells (mDCs) and EPIC does not consider any kind of DCs (supplementary table 2), it is not surprising that both methods view these genes as B cell markers and are not able to discriminate between pDCs and B cells.

Applying the methods to TCGA data reveals reliable signatures

To compare cell type estimates from different methods, we ran the methods on 10,273 primary tumor samples from TCGA[20] and correlated the estimates for each cell type with each other (figure 4). If the estimates of two signatures are correlated, this can be either due to biological reasons, *i.e.* two cell types tend to co-occur; or due to methodological reasons, *i.e.* the signatures are related, picking up the same signal. Expectedly, cell types that are predicted reliably by all methods also form a cluster in the heatmap, suggesting the similarity of the signatures. At the top of the matrix two T-cell clusters emerge, both containing $CD4^+$ and $CD8^+$ T cells, which, together with the spillover analysis, suggests a strong overlap of the signatures for corresponding cell types. This is true in particular for deconvolution methods: unlike marker-based approaches that leverage on a different gene set for each cell type, deconvolution methods always consider all signature genes. We also observed a distinct B-cell cluster, indicating that B cells have a distinct cellular signature, although some genes with lower specificity can hamper its accurate quantification, as discussed in the previous section. Macrophages and monocytes form a cluster at the right bottom of the matrix, suggesting that majority of the methods agree on the estimates of these cell types. Dendritic cells are scattered across the matrix. Only the *dendritic cells resting* signature of CIBERSORT is clustered together with the *dendritic cell* signature of xCell. All other DC signatures appear to be unrelated, again highlighting the heterogeneity of DCs.

Guidelines for method selection

In table 2, we provide guidelines on which method to use for which cell type based on three criteria: interpretability of the score, overall performance and possible limitations. It is very important to understand the implications of the different scoring strategies. EPIC and quanTIseq are the only methods providing an ‘absolute score’ that represents a cell fraction. All other methods provide scores in arbitrary units, that are only meaningful in relation to

another sample of the same dataset. For this reason, and due to a robust overall performance, we recommend EPIC and quanTIseq for general purpose deconvolution. In practice, absolute scores are not always necessary. For instance, in a clinical trial, relative scoring methods can be used to infer fold changes between treatment and control group or to monitor changes of the immune composition in longitudinal samples. In that case, using MCP-counter is a good choice thanks to its highly specific signatures that excelled in the spillover benchmark. A limitation of deconvolution methods is that they are susceptible to background predictions, i.e. prediction of (small) fractions for cell types that are actually absent. Therefore, when interested in presence/absence of a cell type, we suggest using xCell.

Immunodeconv R package provide easy access to deconvolution methods

We created an R package, `immunedeconv`, that provides a unified interface to the different deconvolution methods. The package is freely available from GitHub: https://github.com/grst/immunedeconv. We emphasize the reproducibility and extensibility of our benchmark. A snakemake pipeline[22], including a step-by-step instruction on how to test an additional method is available from https://github.com/grst/immune_deconvolution_benchmark.

Discussion

Quantification of the cellular composition of bulk tumor samples is a challenging computational problem that has been addressed by a variety of methods. For the utility of these methods it is imperative to fully understand their capabilities and limitations. Here we provide a comprehensive overview and the first quantitative comparison of existing computational methods applied to RNA-seq data to guide researchers and clinicians in selecting the most suitable approach for the analysis of their samples.

We assessed seven state-of-art computational methods by applying them to simulated bulk RNA-seq derived from scRNA-seq dataset of 11,000 single cells, integrated from different experiments[18]. ScRNA-seq-based simulation by *in silico* mixing of cell types is a powerful surrogate for bulk-RNA, with the addition that it allows to generate data ad hoc to assess specific performance metrics systematically. Nevertheless, FACS or IHC measurements are currently considered the gold standard for assessing cell type composition and,

unfortunately, in public databases only few RNA-seq samples with matched FACS or IHC reference can be found. At the time of writing, we only managed to get access to three of them[10,18,19]. More such publicly-available matching datasets will likely strongly enhance our ability to reliably estimate correlation of computational deconvolution methods to these gold standard techniques. However, even with limited number of observations available, we showed that there is generally a good correlation between FACS and RNA-seq deconvolution.

The major conclusion of our study is that RNA-seq can be utilized for estimation of immune cell infiltrates in the tumor robustly and with a good accuracy, particularly when a cell type is well characterized. Most methods show near-perfect correlations on $CD8^+$ T-cells, B cells, NK, fibroblasts and endothelial cells. However, as observed previously by Schelker *et al.*[18] in their benchmark of CIBERSORT, regulatory and non-regulatory $CD4^+$ T cells are hard to distinguish. We found that DCs are another example that is difficult to characterize based on the currently available gene signatures. DCs have a variety of subtypes that feature different expression profiles[23,24]. While MCP-counter specifically addresses myeloid dendritic cells (mDCs), all other methods provide general signatures for DCs. However, our results demonstrate that none of methods actually quantify total DCs, but only specific DC subtypes as reported in supplementary table 2. For these “problematic” cell (sub)types, the development of good signatures are essential to improve deconvolution performance and should be the focus of future research. We envision that scRNA-seq data will facilitate these developments by improving our understanding of immune cell types and states, and enabling the definition of more accurate signatures, which then can be later integrated and evaluated using the benchmark presented in this study.

An important outcome of this study is to provide guidelines to method selection, allowing researchers and practitioners to make an informed decision when selecting biomarker strategies for patient biopsies. Deconvolution approaches like EPIC and quanTIseq that generate scores that are interpretable as a cell fraction are promising as they allow for both inter- and intra-sample comparisons and demonstrate robust performance in the benchmark. However, compared to xCell, which uses a statistical test for enrichment of marker genes, they tend to predict small “background” levels.

Our benchmark goes beyond simple method assessment as it provides solutions for some of the found shortcomings. We have demonstrated that the substantial spillover between DCs and B cells observed in several methods was caused by a subset of marker genes characterized by lower cell-specificity. We validated in a public knowledge base that these

genes are indeed markers expressed in both pDCs and B cells on the mRNA level and could significantly reduce the spillover effect by removing the genes from the signature matrices. Similarly, we could substantially reduce the background prediction level of EPIC and quanTIseq on $CD4^+$ T cells by excluding non-genes. Our results demonstrate that explicitly addressing spillover/background effects can lead to a potential improvement of deconvolution methods.

Similar cell type deconvolution methods exist for DNA methylation array or bisulfite sequencing (BS-seq) data[25]. Benchmarking strategy introduced in out study utilizing single-cell sequencing to simulate bulk sequencing would be advantageous to validate such methods. However, while protocols for single cell BS-seq have been established[26], we currently lack a comprehensive single cell methylation data set needed for generating bulk DNA methylation sequencing data as suggested here.

In summary, our results demonstrate that computational deconvolution is feasible at high accuracy, given well-defined, high quality signatures. These findings indicate that future efforts should prioritize defining cell populations and finding reliable signatures over developing new deconvolution approaches. This study establishes a roadmap to evaluate computational cell-type deconvolution methods and provides guidelines to researchers working in immuno-oncology in selecting appropriate methods for characterization of cellular context of the tumors.

Methods

Single-cell data

The dataset of 11,759 single cells from Schelker et al.[18] was obtained through their figshare repository. We reproduced their analysis using their MATLAB source code and exported the final dataset to continue our analysis in R. We excluded all cells that were classified as “Unknown” from the downstream analysis. Moreover, we did not consider cells originating from PBMC as we were interested in the performance of deconvoluting cancer samples. Gene expression values are expressed as transcripts per million (TPM).

Immune-cell reference data

Immune cell reference profiles were obtained from the sources described in supplementary table 3. FASTQ files were extracted from SRA files using the sra toolkit[27] with the following options:

```
>fastq-dump --outdir fastq --gzip --skip-technical --readids --read-filter
pass --dumpbase --split-files --clip
```

Reads were aligned and gene expression estimated as TPM using STAR[28] and RSEM[29] with the following options:

```
>rsem-calculate-expression -p 8 --paired-end --star --star-gzipped-read-
file --no-bam-output --append-names
```

Validation data

We obtained preprocessed bulk RNA-seq and FACS data for 8 PBMC samples from Hoek et al.[19] through personal communication from the quanTIseq authors[9]. We obtained bulk RNA-seq and FACS data for 4 metastatic melanoma samples from Racle et al.[10] through GEO accession number GSE93722. We obtained bulk RNA-seq and FACS data for 3 ovarian cancer ascites samples[18] from the figshare repository and through personal communication. The 3 ovarian cancer ascites samples consist of two technical replicates each, which we merged by computing the arithmetic mean for each gene. All gene expression values are expressed as TPM.

Cell-type mapping

For comparing the methods, it is essential to map the cell types of the different methods to a controlled vocabulary (CV). It has to be taken into account, that the different methods resolve the cell types at different granularities. *E.g.* while CIBERSORT predicts naive- and memory $CD8^+$ T cells, all other methods only predict $CD8^+$ T cells. We address this issue by creating a hierarchy of cell types, and map each cell type from a dataset or method to a node in the cell type tree (see supplementary figure 8 and supplementary table 4). If a node (e.g. $CD8^+$ T cell) is to be compared among the different methods, the estimated fraction is computed as follows: (i) if the method provides an estimate for the node, take that value (ii) if the method provides an estimate for all child nodes, sum up all children. This is done recursively until no estimates are available or the leafs of the tree are reached. (iii) if an estimate is missing for at least one of the child nodes, no estimate will be available. An exception was made for the following cell types, that are only provided by few methods: T cell gamma delta, Macrophage M0, Monocyte, T cell follicular helper. If an estimate is missing for one of those cell types, the remaining child nodes will be summed up.

Validation of the methodology

Schelker et al.[18] provide three ovarian cancer ascites samples for which both single cell RNA-seq and bulk RNA-seq data is available. We generated artificial bulk samples by taking the mean of the TPM values for each gene of all single cell samples. First, we compared the values using Pearson correlation on log scale. Next, we used BioQC[30] to test for differential enrichment of gene ontology (GO)-terms[31]. Finally, we ran the deconvolution methods on both the genuine and simulated bulk RNA-seq sample and compared the results. To demonstrate that xCell's low correlation is in fact due to little variance between the samples, we reran xCell while additionally including the 51 immune cell reference samples (supplementary table 3) in the expression matrix of both simulated and genuine bulk RNA-seq.

Generation of artificial bulk samples

We generated simulated bulk samples from single cell data by calculating the average gene expression of 500 selected cells. Cells were selected randomly as follows: (1) A fraction f of cancer cells was drawn from $\sim N(0.33, 0.3)$ which is the empirical distribution of cancer cells in the melanoma and ascites samples in the single cell dataset. f was constrained to the interval [0, 0.99].
(2) Half of the samples use melanoma cancer cells, the other half ovarian cancer cells
(3) The remaining fraction $1 - f$ was randomly assigned to immune cells, resulting in a fraction vector
(4) The fraction vector was multiplied with 500 to obtain a cell count vector
(5) The corresponding number of cells was drawn with replacement from the single cell dataset. That implies that for cell populations with only few cells available, the artificial bulk sample will contain the same single cell sample multiple times.

Deconvolution methods

We implemented an R package, `immunedeconv`, that provides a unified interface to all seven deconvolution methods compared in this paper. The CIBERSORT R source code was obtained from their website on 2018-03-26. The xCell, EPIC, MCP-counter, TIMER, quanTIseq source codes were obtained from GitHub from the following commits: dviraran/xCell@870ddc39, GfellerLab/EPIC@e5ae8803, ebecht/MCPcounter@e7c379b4, hanfeisun/TIMER@a030bac73, FFinotello/quanTIseq@ee9f4036.

We ran CIBERSORT with disabled quantile normalization, as recommended on their website for RNA-seq data. While for consistency, quanTIseq provides an entire pipeline, starting with read mapping and estimation of gene expression, we only ran the last part of that pipeline, which estimates the immune cell fractions from gene expression data. We ran TIMER with "OV" on ovarian cancer ascites samples and with "SKCM" on melanoma samples. We ran quantiseq with the option `tumor = TRUE` on all tumor samples and `tumor = FALSE` on the PBMC samples. We ran EPIC with the `TRef` signature set on all tumor samples and with `BRef` on the PBMC samples. We disabled the mRNA scaling options of quanTIseq and EPIC for the single cell simulation benchmark using the `mRNAscale` and `mRNA_cell` options, respectively. Notably, this only has an effect on the results of the absolute quantification, but not on the correlations used to compare all methods. For each of the datasets (simulated, Hoek, Schelker, Racle), we submitted all samples in a single run.

TCGA data

We downloaded the 2016-01-28 analysis run of TCGA gene expression from firebrowse.org[32]. We obtained TPM from the `rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes__data.data.txt` file by taking the `scaled_estimate` value and multiplying it with 1e6.

Simulation benchmark

Known fractions of simulated samples are compared to the predicted scores using Pearson's correlation coefficient *r*. P-values and *r* have been computed using the `cor` function in R.

Minimal detection and background prediction levels

We generated simulated bulk samples by taking the arithmetic mean of single cell gene expression values. 1,806 non-immune cells (fibroblasts, endothelial cells, ovarian cancer cells and melanoma cells) were used as a background. For each type of immune cell an increasing number of this cell type (0, 5, 10, … 50, 60, 70, … 100, 150, 200, … 500, 600, … 1000) was sampled randomly from the pool of single cells and added to the background. For each number of immune cells 5 replicates were sampled to obtain a confidence interval, leading to five batches of 870 samples (145 spike-in levels x 6 immune cell types). Each of the five batches was submitted to the methods independently in a single run. We quantify the detection limit as the minimal fraction of added immune cell of the corresponding type at which the predicted fraction is significantly different from 0 (one-sided t-test, alpha = 0.05).

We quantify false-positive predictions as the predicted fraction of cells of a certain type when there are actually none, *i.e.* only background cells.

Spillover

Analogous to FACS, we defined spillover as the prediction of another cell type due to a partial overlap of the signatures. We measured spillover on three datasets: (1) 44 reference samples (supplementary table 3) of pure immune cells. (2) 30 simulated samples of pure immune cells of six types (3) 30 simulated samples consisting of 100 immune cells of six types and 400 non-immune cells as background. Each dataset was submitted to the methods independently in a single run. We visualized the results as chord diagrams (figure 3a) using the R package `circlize`[33].

Other tools

Reproducible reports were generated using bookdown 0.7[34]. We make use of Snakemake 5.2.0[22], conda and bioconda[35] to integrate our analyses in a reproducible pipeline.

Code availability

An R package providing a unified interface to the deconvolution methods is freely available from GitHub: https://github.com/grst/immunedeconv. A Snakemake pipeline[22] to reproduce the results is available from https://github.com/grst/immune_deconvolution_benchmark.

Data availability

Preprocessed data are available from GitHub along with all code necessary to reproduce our results: https://github.com/grst/immune_deconvolution_benchmark. Single cell and ovarian cancer ascites data has been downloaded from https://figshare.com/s/711d3fb2bd3288c8483a. The immune cell reference profiles have been obtained from the sources listed in supplementary table 3.

Table 1: Overview of methods considered in this benchmark. Methods can be conceptually distinguished in marker-gene-based approaches (M) and deconvolution-based approaches (D). The output scores of the methods have different properties and allow either intra-sample comparisons between cell types, inter-sample comparisons of the same cell type, or both. All methods come with a set of cell type signatures ranging from 6 immune cell types to 64 immune and non-immune cell types.

tool	abbreviation	approach	score	comparisons	cell types	reference
CIBERSORT	CBS	D	immune cell fractions, relative to total immune cell content	intra	22 immune cell types	[8]
CIBERSORT absolute mode	CBA	D	Score of arbitrary units that reflects the absolute proportion of each cell type	intra, inter	22 immune cell types	[8]
EPIC	EPC	D	Cell fractions, relative to all cells in sample	intra, inter	6 immune cell types, fibroblasts, endothelial cells	[10]
MCP-counter	MCP	M	Arbitrary units, comparable between samples	inter	8 immune cell types, fibroblasts, endothelial cells	[5]
quanTIseq	QTS	D	Cell fractions, relative to all cells in sample	intra, inter	10 immune cell types	[9]
TIMER	TMR	D	Arbitrary units,	inter	6 immune cell	[11]

			comparable between samples (not different cancer types)		types	
xCell	XCL	M	Arbitrary units, comparable between samples	inter	64 immune and non-immune cell types	[6]

Table 2: Guidelines for method selection. For each cell type, we recommend which method to use, highlighting advantages and possible limitations. *Overall performance*: Indicates how well predicted fractions correspond to known fractions in the benchmark. Absolute score: the method provides an absolute score that can be interpreted as a cell fraction. Methods that do not support an absolute score only supports inter-sample comparison within one experimental dataset, *i.e.* the score is only meaningful in relation to another sample f. *Background predictions*: Indicates, if a method tends to predict a cell type although it is absent.

Cell type	**Recommended methods**	**Overall performance**	**Absolute score**	**No background predictions**
B	EPIC	++	++	+
	MCP-counter	++	-	-
T CD4$^+$	EPIC	++	++	-
	xCell	++	-	++
T CD4$^+$ non-reg.	quanTIseq	+	++	-
	xCell	+	-	++
T reg.	quanTIseq	++	++	-
	xCell	++	-	++
T CD8$^+$	quanTIseq	++	++	+
	EPIC	++	++	-
	MCP-counter	++	-	-
NK	EPIC	++	++	+
	MCP-counter	++	-	-
Macrophages/Monocytes	EPIC	++	++	+
	MCP-counter	++	-	-
DC	None of the methods can be recommended to estimate overall DC content. MCP-counter and quanTIseq can be used to profile myeloid DCs.			

Figure 1: (a) Results of the benchmark on 100 simulated samples. The plots show the correlation of predicted vs. known cell type fractions on artificial bulk RNA-seq samples generated from single cell RNA-seq. Pearson's *r* is indicated in each subplot. Due to the lack of a corresponding signature, we estimated macrophages/monocytes with EPIC using the "macrophage" signature and with MCP-counter using the "monocytic lineage" signature as a surrogate. (b) Performance of the methods on three independent datasets that provide immune cell quantification by FACS. Different cell types are indicated in different colors. Pearson's *r* has been computed as a single correlation on all cell types simultaneously. Note that only methods that allow both inter- and intra-sample comparisons (*i.e.* EPIC, quanTIseq, CIBERSORT absolute mode) can be expected to perform well here. (c-d) Performance on the three validation datasets per cell type. Schelker's and Hoek's dataset have too few samples to be considered individually. The values indicate Pearson correlation of the predictions with the cell type fractions determined using FACS. Blank squares indicate that the method does not provide a signature for the respective cell type. 'n/a' values indicate that no correlation could be computed because all predictions were zero. The asterisk (*) indicates that the 'monocytic lineage' signature was used as a surrogate to predict monocyte content. P-values: **** < 0.0001; *** < 0.001; ** < 0.01; * < 0.05; ns ≥ 0.05. P-values are not adjusted for multiple testing. Method abbreviations: see table 1.

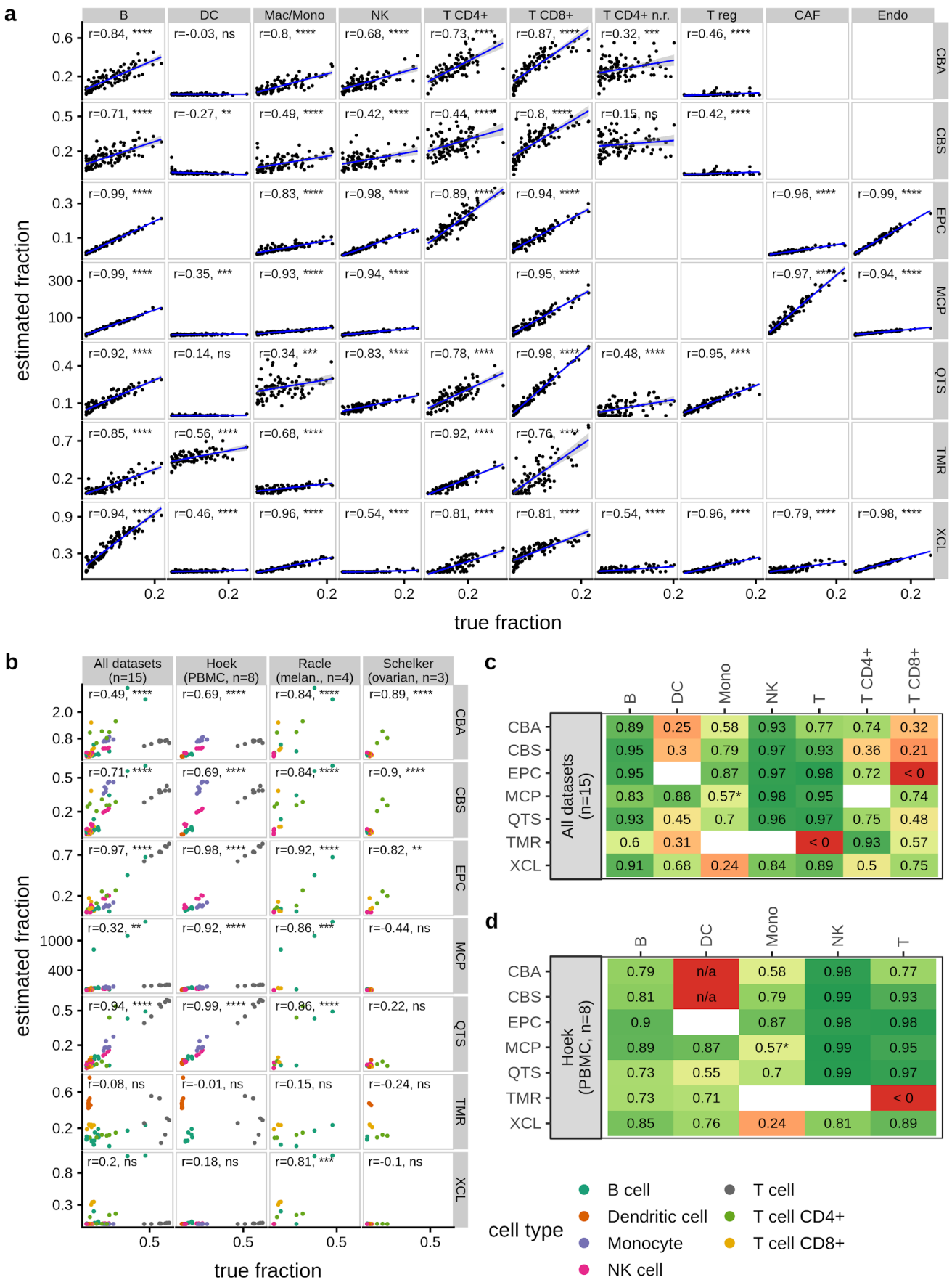

a
B
DC
Mac/Mono
NK
T CD4+
T CD8+
T CD4+ n.r.
T reg
CAF
Endo
CBA
CBS
EPC
MCP
QTS
TMR
XCL
estimated fraction
true fraction
b
All datasets (n=15)
Hoek (PBMC, n=8)
Racle (melan., n=4)
Schelker (ovarian, n=3)
c
All datasets (n=15)
B
DC
Mono
NK
T
T CD4+
T CD8+
d
Hoek (PBMC, n=8)
cell type
B cell
Dendritic cell
Monocyte
NK cell
T cell
T cell CD4+
T cell CD8+

Figure 2: Minimal detection fraction and background prediction level. The plots show the average estimate of each method for five simulated bulk RNA-seq datasets randomly sampled from ~1,800 non-immune cells and an increasing fraction of a specific immune cell type. The grey ribbon indicates the 95% confidence interval. The red line refers to the *minimal detection* fraction, *i.e.* the minimal fraction of an immune cell type needed for a method to reliably detect its abundance as significantly different from the background (p-value < 0.05, one-sided t-test). The blue line refers to the *background prediction* level, *i.e.* the average estimate of a method with no immune cells present.

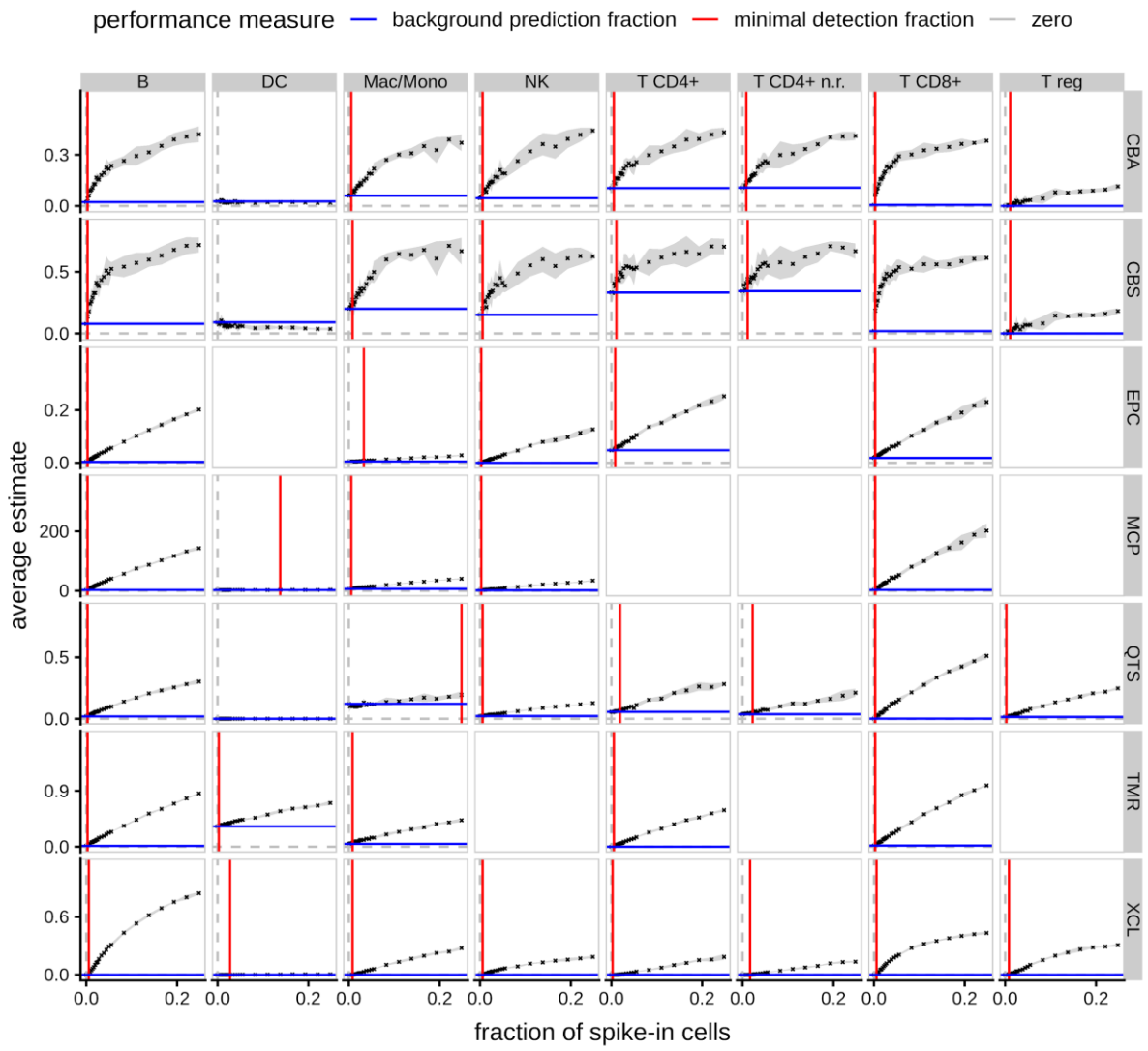


Figure 3: Spillover analysis. All methods were applied to 30 simulated bulk RNA-seq samples of six immune cell types. (a) The plots show the ratio of correct vs. false prediction and unveil colinearities between the cell types. CIBERSORT shows identical results in both standard and absolute mode in this analysis. $CD4^+$ T cell samples are an aggregate of regulatory and non-regulatory $CD4^+$ T cells. The numbers indicate the overall noise ratio, *i.e.* the fraction of predictions that are attributed to a wrong cell type. (b) Predicted $CD4^+$ T cell score on ten simulated bulk RNA-seq samples that only contain non-immune cells (fibroblast, endothelial and cancer cells) before and after removing five signature genes that we found to be specific for both $CD4^+$ T cells and the non-immune cell populations. (c) Predicted B cell content on ten simulated bulk RNA-seq samples that only contain plasmacytoid dendritic cells (pDCs) before and after removing six signature genes that we found to be specific for both pDCs and B cells. P-values were computed using paired two-sided t-test with 9 degrees of freedom and have not been adjusted for multiple testing. The black horizontal bar shows the mean of the data points.

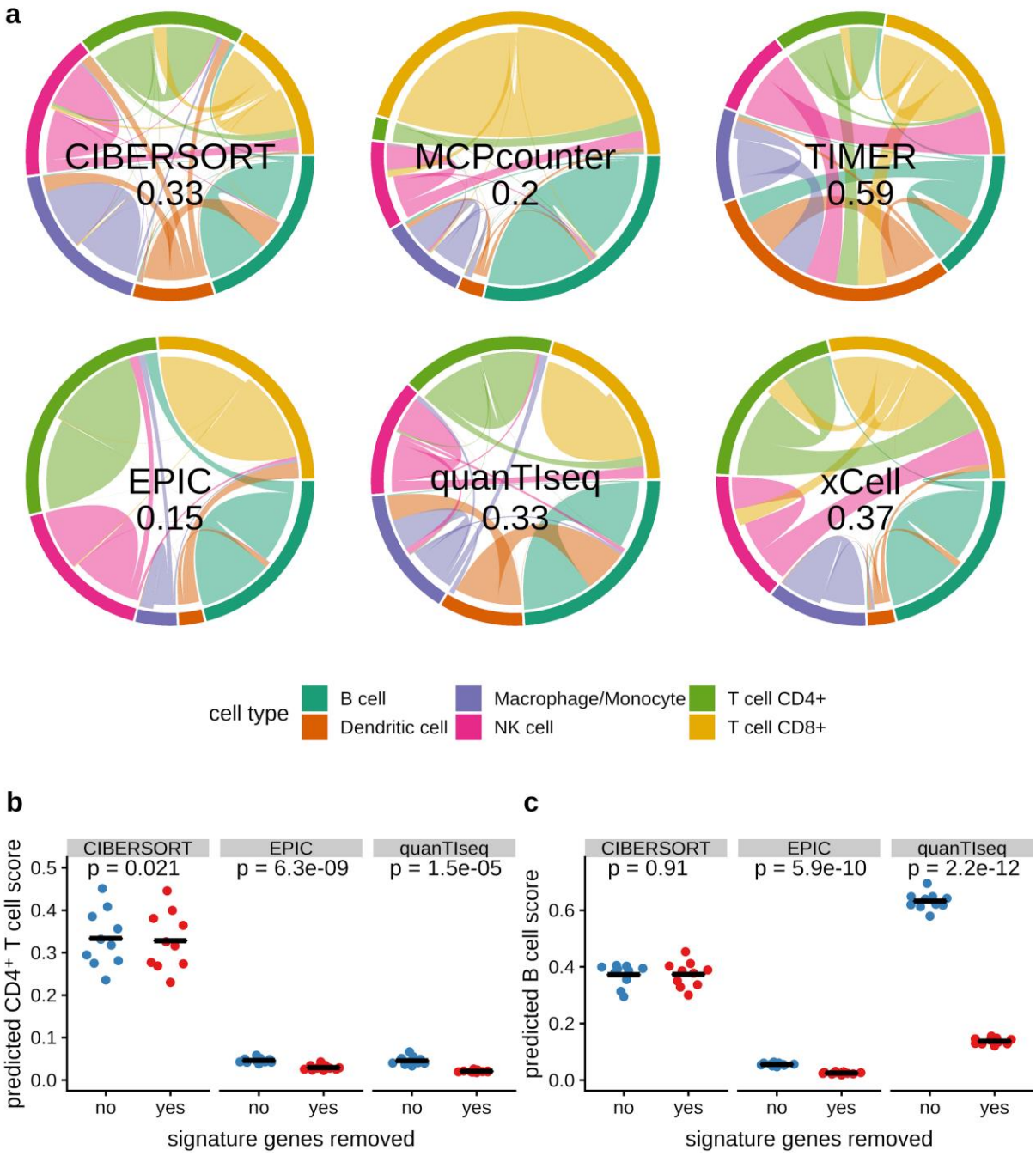
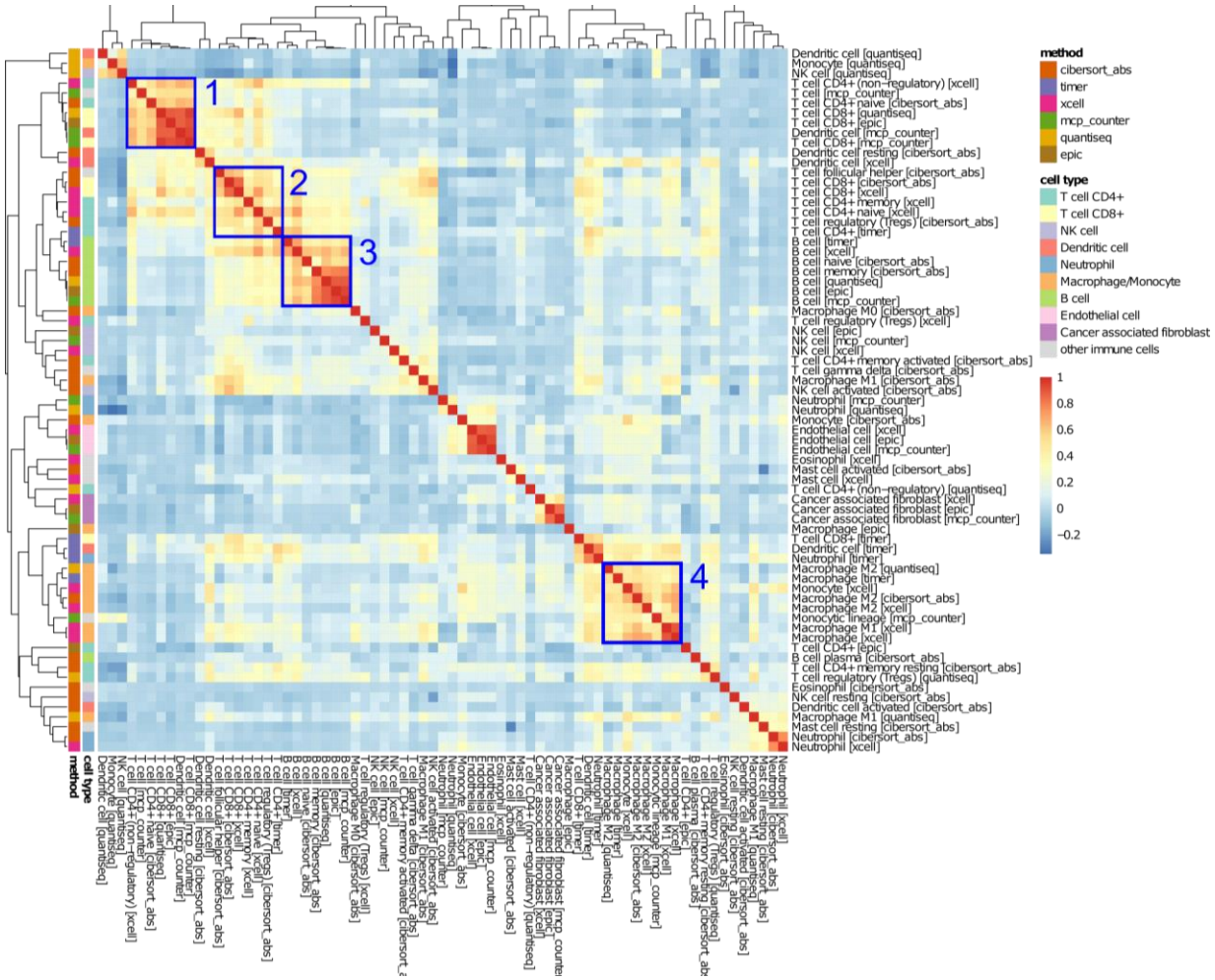

a
CIBERSORT
0.33
MCPcounter
0.2
TIMER
0.59
EPIC
0.15
quanTIseq
0.33
xCell
0.37
cell type
B cell
Dendritic cell
Macrophage/Monocyte
NK cell
T cell CD4+
T cell CD8+
b
CIBERSORT
p = 0.021
EPIC
p = 6.3e-09
quanTIseq
p = 1.5e-05
predicted CD4+ T cell score
0.0
0.1
0.2
0.3
0.4
0.5
no
yes
signature genes removed
c
CIBERSORT
p = 0.91
EPIC
p = 5.9e-10
quanTIseq
p = 2.2e-12
predicted B cell score
0.0
0.2
0.4
0.6
signature genes removed

Figure 4: Correlations of signatures. All methods were applied to 10,273 samples from TCGA. The matrix show the Pearson correlation of the predictions of the different signatures of the different methods. Both rows and columns are clustered hierarchically by Pearson correlation. The clusters highlighted in the figure consist of T cell signatures (1 and 2), B cell signatures (3) and Macrophage/Monocyte signatures (4).



Declarations

Availability of data and material

All data and software needed to replicate the study are available on GitHub (grst/immune_deconvolution_benchmark).

Competing interests

Both GS and TA are current employees of Pieris Pharmaceuticals GmbH, Germany. JDZ is current employee of F. Hoffmann-La Roche Ltd., Switzerland.

Authors' contributions

Study design: GS, JB, WHF and TA. Data analysis: GS and JDZ. Writing manuscript: GS, FF, FP, ML and TA. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Max Schelker for sharing additional data and being very responsive to questions regarding their single cell dataset. We thank Hubert Hackl for suggesting to mine TCGA data to assess the heterogeneity of signatures.

References

1. Fridman, W. H., Pagès, F., Sautès-Fridman, C. & Galon, J. The immune contexture in human tumours: impact on clinical outcome. *Nat. Rev. Cancer* **12,** 298–306 (2012).
2. Friedman, A. A., Letai, A., Fisher, D. E. & Flaherty, K. T. Precision medicine for cancer with next-generation functional diagnostics. *Nat. Rev. Cancer* **15,** 747–756 (2015).
3. Petitprez, F., Sun, C. M. & Lacroix, L. Quantitative analyses of the tumor microenvironment composition and orientation in the era of precision medicine. *Frontiers in* (2018). doi:10.3389/fonc.2018.00390
4. Lambrechts, D. *et al.* Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat. Med.* **24,** 1277–1289 (2018).
5. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17,** 218 (2016).
6. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18,** 220 (2017).
7. Finotello, F. & Trajanoski, Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. *Cancer Immunol. Immunother.* **0,** (2018).
8. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 453 (2015).
9. Finotello, F. *et al.* quanTIseq: quantifying immune contexture of human tumors. *bioRxiv* 223180 (2017). doi:10.1101/223180
10. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife* **6,** e26476 (2017).
11. Li, B. *et al.* Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* **17,** 174 (2016).
12. Newman, A. M. & Alizadeh, A. A. High-throughput genomic profiling of tumor-infiltrating leukocytes. *Curr. Opin. Immunol.* **41,** 77–84 (2016).

13. Avila Cobos, F., Vandesompele, J., Mestdagh, P. & De Preter, K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* **34,** 1969–1979 (2018).
14. Petitprez, F. *et al.* Transcriptomic analysis of the tumor microenvironment to guide prognosis and immunotherapies. *Cancer Immunol. Immunother.* **67,** 981–988 (2018).
15. Newman, Aaron M Liu, Chih Long Green, Michael R Gentles, Andrew J Feng, Weiguo Xu, Yue Hoang, Chuong D Diehn, Maximilian Alizadeh, Ash A. CIBERSORT website. *CIBERSORT* Available at: https://cibersort.stanford.edu/. (Accessed: 20th October 2018)
16. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007).
17. Aran, D. xCell repository on GitHub. (2018). Available at: https://github.com/dviraran/xCell/blob/ce4d43121c4a161b1e72a50dc875e43d9cf89b0d/README.Md. (Accessed: 20th September 2018)
18. Schelker, M. *et al.* Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.* **8,** 2032 (2017).
19. Hoek, K. L. *et al.* A Cell-Based Systems Biology Assessment of Human Blood to Monitor Immune Responses after Influenza Vaccination. *PLoS One* **10,** e0118528 (2015).
20. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–1120 (2013).
21. Edgar, R. *et al.* LifeMap Discovery™: the embryonic development, stem cells, and regenerative medicine research portal. *PLoS One* **8,** e66629 (2013).
22. Köster, J. & Rahmann, S. Building and documenting workflows with python-based snakemake. in *OASIcs-OpenAccess Series in Informatics* **26,** (Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012).
23. Collin, M., McGovern, N. & Haniffa, M. Human dendritic cell subsets. *Immunology* **140,** 22–30 (2013).

24. Villani, A.-C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356,** (2017).
25. Teschendorff, A. E., Breeze, C. E., Zheng, S. C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics* **18,** 105 (2017).
26. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11,** 817–820 (2014).
27. Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* **39,** D19–21 (2011).
28. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
29. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).
30. Zhang, J. D. *et al.* Detect tissue heterogeneity in gene expression data with BioQC. *BMC Genomics* **18,** 277 (2017).
31. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45,** D331–D338 (2017).
32. Broad Institute TCGA Genome Data Analysis Center. Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run. (2016). doi:10.7908/C11G0KM9
33. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in R. *Bioinformatics* **30,** 2811–2812 (2014).
34. Xie, Y. *bookdown: Authoring Books and Technical Documents with R Markdown*. (CRC Press, 2016).
35. Grüning, B. *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* **15,** 475–476 (2018).