multideconv **-** an integrative pipeline for combining first and second generation cell type deconvolution results

Marcelo Hurtado[1,2], Abdelmounim Essabbar[1,2], Leila Khajavi[1†], Vera Pancaldi[1,2]

[1]Univ Toulouse, CNRS, INSERM, CRCT, Toulouse, France

[2] Equipe labellisée Ligue Contre le Cancer

† Present address: Bioinformatics Department, Evotec, Toulouse, France

Abstract

Summary: The number of computational methods for cell type deconvolution from bulk RNA-seq data has been increasing in the last years, but their high feature complexity and variability of results across methods and signatures limit their utility and effectiveness for patient stratification. Applying multiple combinations of deconvolution methods and signatures often results in hundreds of redundant or contradictory cell type features describing the composition of complex tumour samples. Benchmarking efforts are inherently limited by the lack of bias-free ground truth, often yielding inconsistent results or no consensus. To address these limitations, we present *multideconv*, an R package that reduces dimensionality and eliminates redundancy in deconvolution results, through unsupervised filtering and iterative correlation analyses. Built on top of existing frameworks, *multideconv* harmonizes outputs across methods to identify robust cell type proportion estimates and mitigate signature-driven heterogeneity.

Availability and implementation: The *multideconv* R package and tutorials are available at ***https://github.com/VeraPancaldiLab/multideconv***. The code to reproduce the analysis and figures is available on github at ***https://github.com/VeraPancaldiLab/multideconv_paper***. The Mariathasan et al. datasets used to showcase the method can be found in the IMVigor210Biologies R Package. The Gide et al. bulk RNA dataset is available in the European Nucleotide Archive (ENA) under accession number PRJEB23709. The scRNAseq LUAD cohort (Vanderbilt) data is available on Zenodo under accession number 7878082.

Contact: marcelo.hurtado@inserm.fr or vera.pancaldi@inserm.fr

Supplementary information: Supplementary materials are available at Supplementary materials - multideconv

Introduction

Inferring cell type composition from complex samples, such as blood and other tissues, and its evolution across time or in disease has relevance across several biomedical fields as well as in developmental biology, but especially in immuno-oncology. The tumor microenvironment (TME), composed of immune, stromal, and other cells, plays a key role in modulating tumor progression. Despite significant progress in understanding this complex system, it remains

unclear why some patients respond to specific therapies while others do not and experience recurrence. Accurately characterizing the TME is critical for understanding tumor behavior and supporting treatment decisions (Hurtado et al. 2024). Current technologies, such as flow cytometry, immunohistochemistry (IHC), or single-cell RNA sequencing (scRNA-seq), provide high-resolution cellular phenotyping. However, inherent biases due to cell isolation, their high cost and technical complexity, as well as data interpretation challenges, limit their routine application in clinical settings. As a result, computational methods for estimating cell type composition from more affordable and practical bulk transcriptomics experiments have continued to gain interest in the last few years.

Reference-based deconvolution methods are commonly categorized into first- and second-generation approaches. While first-generation methods rely on static gene expression signatures derived from purified cell populations, second-generation methods leverage single-cell RNA-seq data to construct dynamic, context-specific signatures that can model a broader spectrum of cell types and states. Although more flexible, second-generation methods also introduce new challenges, including variability in single-cell references, biases in cell type annotations, inconsistent marker selection, and increased computational cost.

Recent initiatives like immunedeconv (Sturm et al. 2019) and omnideconv (Dietrich et al. 2024, Vathrakokoili et al. 2024) have attempted to standardize and benchmark these tools, providing a single point of access to several first- and second-generation deconvolution methods for easy application to gene expression datasets. However, due to the absence of absolute bias-free ground truth, benchmarking efforts often produce discordant or context-specific outcomes, making it difficult to generalize results or define a universally "best" method. Some approaches have sought to reconcile method differences by deriving a consensus output (Jiménez-Sánchez et al. 2019), but this can oversimplify the biological complexity and fail to exploit the complementary information provided by individual tools. To capture the full complexity of the TME, it is often necessary to apply multiple combinations of methods and signatures. While this approach broadens biological insight, it also results in high-dimensional and heterogenous features that hinder interpretation and complicate downstream analysis.

To address these issues, we introduce *multideconv*, an R package designed to reduce the dimensionality and redundancy of deconvolution outputs. *multideconv* integrates the outputs of first- and second-generation methods and employs a combination of unsupervised filtering, linear correlation analysis, and proportionality-based measures to identify robust cell type features across samples. By systematically harmonizing and refining TME deconvolution features, *multideconv* enhances interpretability and facilitates patient stratification across cohorts. Built on top of existing deconvolution ecosystems and compatible with any deconvolution method, *multideconv* offers a streamlined solution for bringing coherence and interpretability in the estimate of complex tissue sample composition ***(Figure 1)***.

Methods

multideconv considers as input RNAseq raw counts with gene symbols, it performs TPM normalization followed by cell type deconvolution. By default, *multideconv* uses six deconvolution methods (CIBERSORTx, DWLS, MOMF,

QuanTIseq, EpiDISH and DeconRNASeq) (***Supplementary Table 1***) combined with nine different cell signatures based on gene expression, methylation data and single cell (***Supplementary Table 2***). If scRNAseq is provided, *multideconv* also includes second-generation methods that require single cell data (MuSic, SCDC, BayesPrism, AutogeneS, CPM and Bisque) (***Supplementary Table 1***). Additionally, *multideconv* provides a function to create gene signatures using four methods (DWLS, BSeq-SC, CIBERSORTx and MOMF) (***Supplementary Table 1***) from the single cell object. These signatures will be used as additional input matrices for the six default methods.

To process the deconvolution features, *multideconv* applies a combination of unsupervised filtering techniques and iterative linear based correlations to form subgroups across all cell types (***Supplementary Figure 1***). The algorithm returns a simplified deconvolution matrix composed with subgroups of results of specific method-signature combinations based on their similarity across samples.

The first step involves unsupervised preprocessing of the deconvolution matrix. Missing values are replaced with zeros to ensure data consistency, while features with high zero occurrences (e.g., >90%) are removed, as these could bias subsequent analyses. Additionally, features with low variance across samples are filtered out, as they likely lack biological significance.

Following preprocessing, the deconvolution matrix is split into cell-type-specific groups based on established naming conventions (e.g., macrophages, NK cells, CD4/CD8 T cells) see ***Supplementary Table 3***, ensuring that features are analyzed within biologically relevant categories. Within each cell-type group, highly correlated features (Pearson's $r > 0.9$) are identified and pruned to reduce redundancy, retaining only one representative feature from each correlated cluster..

Subsequently, features within each cell type are further divided into subgroups using two complementary criteria. First, a proportionality-based correlation analysis is performed to group features with similar proportional changes across samples, after which the median value of each subgroup is computed. Second, a linear-based correlation approach is used to group features with high linear correlation (Pearson's $r > 0.7$) iteratively, and again, their median values are used as representative features. Finally, the resulting subgroups are inspected to remove those derived exclusively from the same deconvolution method, ensuring that the retained subgroups reflect diverse methodological inputs and not method bias. Each subgroup is labeled with its contributing features and methods to support downstream interpretation.

As *multideconv* includes second-generation methods that allow users to input single-cell RNA-seq data and construct cell-type signatures, these objects are often high-dimensional, sparse, and computationally demanding. To address this, *multideconv* implements a metacell construction step per cell type and patient using a k-nearest neighbors (KNN) algorithm adapted from the R package hdWGCNA (Morabito et al. 2023, Langfelder et al. 2008), effectively reducing the complexity of the input single-cell matrix. This task is parallelized across clusters using N workers defined by the user, significantly improving computational efficiency. Once the reduced single-cell object is built, deconvolution is performed using this compact representation, allowing for scalable integration of second-generation data while maintaining computational efficiency and performance (***Figure 1***).

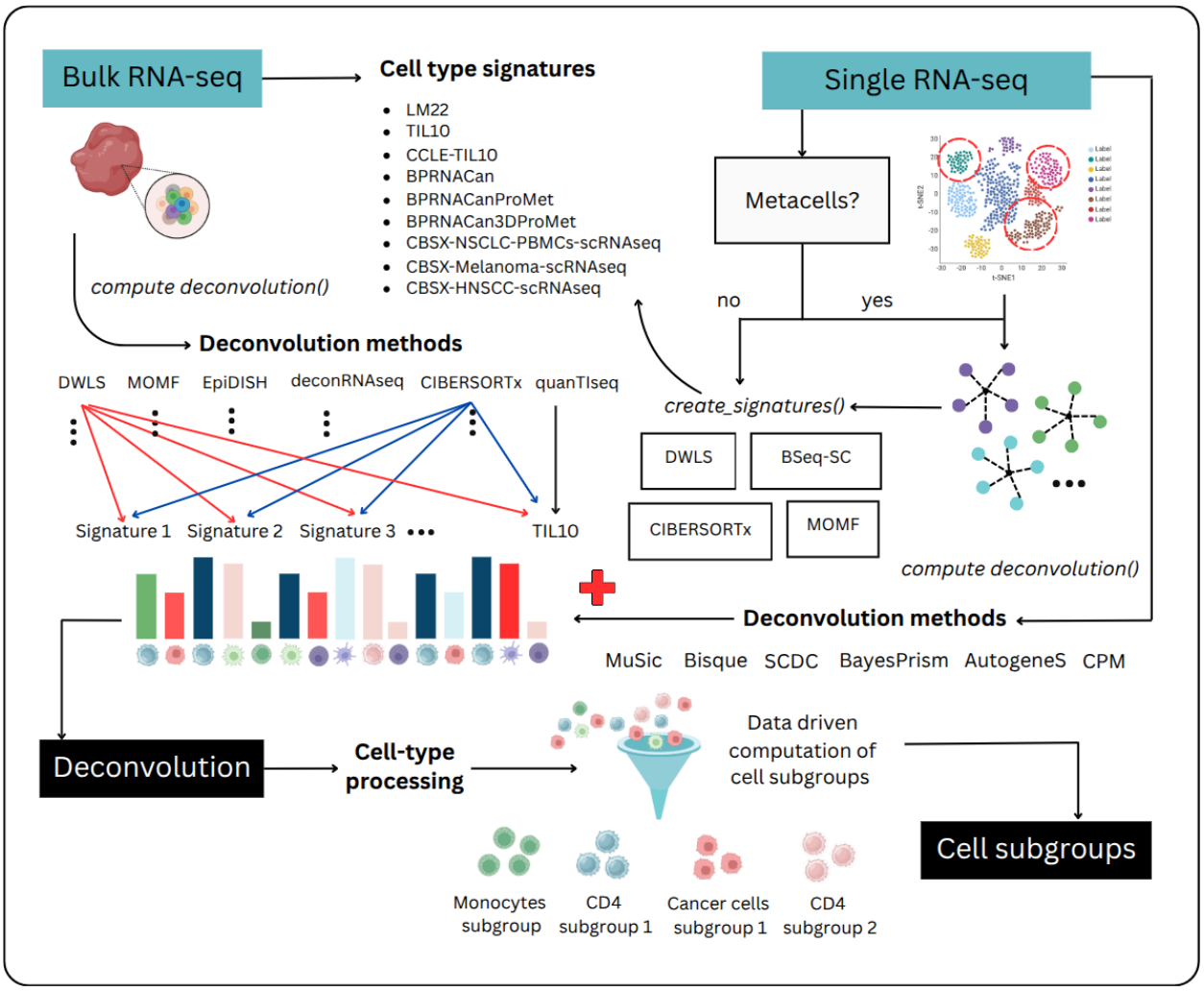

Figure 1. A schematic overview of the *multideconv* pipeline

Results

We use *multideconv* to compute deconvolution of bulk RNA-seq data from bladder cancer samples (Mariathasan et al. 2018), producing 399 deconvolution features using the default methods and signatures. After subgrouping, we are left with 114 grouped deconvolution features.

Grouped features preserve sample clustering structure

We assessed the quality of the cell-type subgroups by comparing their silhouette scores to those derived from an unsupervised clustering approach. For each cell type with more than ten associated features, we extracted the corresponding subset from the original deconvolution matrix and computed a correlation-based distance matrix. Hierarchical clustering was performed on this matrix, and the optimal number of clusters was determined using the gap statistic. The resulting clusters served as a baseline, and their quality was quantified using silhouette scores. We

then evaluated the silhouette scores of our cell-type subgroups, computed using the same distance metric. This comparison allowed us to quantitatively assess how well our predefined clusters capture structure in the data relative to a data-driven clustering baseline. Scores are summarized in ***Table 1,*** where we can see the overall silhouette scores for the subgroups are higher than the baseline, suggesting not only better cohesion, but also that grouped features may even capture more coherent biological structure than unsupervised methods.

Table 1. Silhouette scores

Cell types	Raw deconvolution	Subgroupped features
B cells	0.042	0.455
Macrophages M1	0.035	0.385
Macrophages M2	0.037	0.278
Monocytes	0.044	0.539
Neutrophils	-0.02	0.567
CD4 cells	0.006	0.089

Grouping features preserves original data structure

To assess whether the algorithm preserves the underlying structure of the data after subgrouping, we performed principal component analysis (PCA) on both the original deconvolution matrix and the subgrouped feature matrix. Prior to PCA, we filtered out low-variance and low abundance features to retain informative signals. Visual comparison of PCA projections showed similar distributions between raw and subgrouped features ***(Supplementary Figure 2)***, and the variance explained by the first two principal components was comparable in both cases (raw = 20% PC1 and 11% PC2, subgrouped = 14% PC1 and 10% PC2). Additionally, k-means clustering (k = 3) applied to the top ten principal components from each dataset produced similar average silhouette scores, indicating that the subgrouping strategy maintains the structural integrity of the original data (raw = 0.0897, subgrouped = 0.1545).

Grouped features are highly predictive of immunotherapy response

Deconvolution features can be used to train machine learning models and predict clinical variables. In this line, we decided to use a bulk RNAseq dataset from a melanoma cohort (Gide et al. 2019), for a set of 73 patients with known response to immunotherapy, and assess whether there was a difference in performance between using the raw deconvolution features and the subgrouped ones. Machine learning models were trained using a repeated and stratified k-fold cross-validation (k = 5, repeats = 20), where hyperparameter tuning was performed within the cross-validation loop using grid search, optimizing the area under the ROC curve (AUROC) values. Raw model outputs were calibrated using Platt's logistic model: out-of-fold predictions from cross-validation were used to fit a logistic regression model, which was then applied to transform the raw probabilities into calibrated scores. We benchmarked a diverse set of classification algorithms using the caret R package, spanning tree-based methods

(treebag, rf, C5.0, rpart, xgbTree), linear models (glm, glmnet for LASSO and RIDGE), support vector machines (svmRadial, svmLinear), instance-based learning (knn), and linear discriminant analysis (lda).

Results showed that overall performance is similar whether using raw deconvolution features or subgroups (***Figure 2***). In order to evaluate the feature importance we used SHAP values, which provide a measure of the contribution of each feature to the model's predictions. For each model in the ensemble, the data was split into training and test folds using k-fold cross-validation. The model was trained on the training set, and SHAP values were computed for the test set using the kernelshap method, with the training set as the background data for the SHAP calculation. The SHAP values for each fold were aggregated by averaging across all test samples, providing a feature importance score for each variable. These values were then averaged across all cross-validation folds to obtain a robust estimate of feature importance. As observed in ***Supplementary Figure 3***, subgroups retain high SHAP values for model predictions. These results suggest that our subgrouping approach retains biological variation and produces predictive variables, while offering a more interpretable and compact representation. The machine learning model performance demonstrated that similar results can be achieved using this reduced approach with fewer features, thereby improving computational efficiency and reducing computational time.

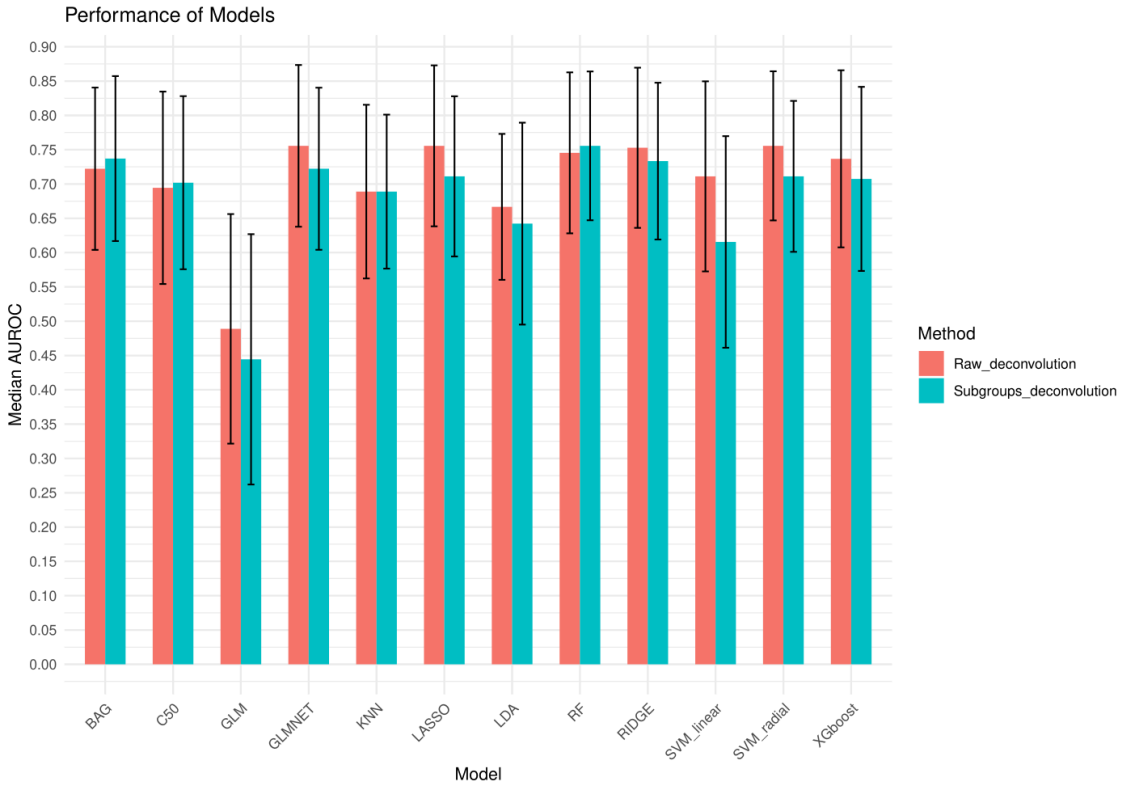

Figure 2. Machine learning model performance. AUROC values using raw 399 deconvolution features (red) and 114 deconvolution subgroups (blue). The bar height corresponds to the AUC-ROC metric averaged across 5 cross-validation schemes (k = 5) with 20 repeats, and the error bar corresponds to ± 1 standard deviation, estimated across the 100 cross-validation schemes.

Validating subgrouped features using scRNAseq datasets

To exemplify the application of *multideconv* on scRNA-seq data, we used a lung cancer dataset including 15 patients (Senosain et al. 2023) ***(Supplementary Figure 4).*** We first computed metacells ***(Supplementary Figure 5)*** and used the reduced single cell object to create static signatures before performing first- and second-generation deconvolution. To evaluate performance, we calculated pseudo-bulk profiles in the original single cell objects, essentially simulating a bulk RNAseq dataset, and deconvolved this pseudo-bulk using the specifically generated single-cell signatures and all the deconvolution methods. This procedure produced a total of 689 features and, as shown in ***Figure 3***, comparison of the deconvolution results and the real cell proportions gave an overall good result for the new signatures (DWLS r-corr = 0.95, CBSX r-corr = 0.91, MOMF r-corr = 0.9 and BSeqSC r-corr = 0.85) suggesting no loss of information during the creation and use of the reduced object for creating the signatures.

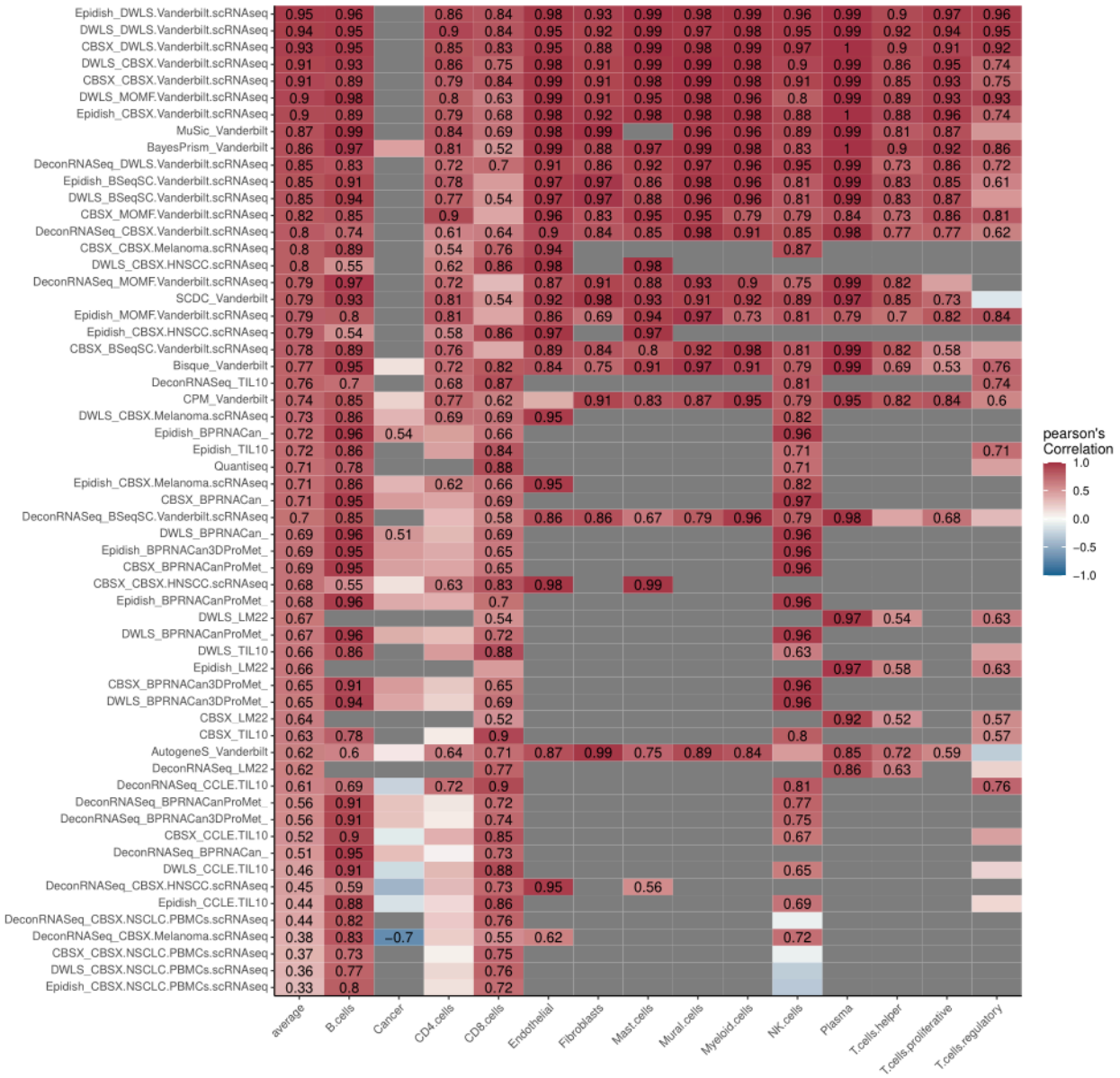

Figure 3. Performance of different methods and signature combinations on the pseudo bulk RNAseq dataset from 15 NSCLC samples for which cell type proportion (ground truth) was estimated using the LUAD scRNAseq on the same samples (Senosain et al. 2023). Pearson's correlations are shown as colour from -1 (blue) to +1 (red) only when p val<0.05. Non-significant correlations are left unlabeled. Grey boxes indicate cell types that were not estimated because the corresponding signature does not include that cell type. Rows indicate different deconvolution features, columns show the different cell types identified.

Conclusion

multideconv offers a practical and scalable solution for navigating the growing complexity of mixed samples deconvolution. By integrating both first- and second-generation deconvolution methods with a diverse set of static and dynamic cell-type signatures, *multideconv* transforms high-dimensional, heterogeneous features into a more biologically meaningful representation. These refined cell-type subgroups are not only easier to manage and interpret, but also better suited for downstream analyses. As the need for integrative and scalable tools continues to grow, *multideconv* offers a modular pipeline that can be easily adapted to more methods and signatures.

Author contributions

M.H. and V.P. conceived the project. M.H. developed the pipeline and performed the analyses. L.K. and A.E. tested and provided valuable discussion on the pipeline methods. V.P. and M.H wrote and reviewed the manuscript. V.P. supervised the project.

Acknowledgements

The authors thank Hafida Hamdache and Francesco Massaini for testing the pipeline and providing useful feedback.

Funding

Work in the Pancaldi lab was funded by the Chair of Bioinformatics in Oncology of the CRCT (INSERM; Fondation Toulouse Cancer Santé and Pierre Fabre Research Institute) and Ligue Nationale Contre le Cancer. This study has been partially supported through the grant EUR CARe N°ANR-18-EURE-0003 in the framework of the Programme des Investissements d'Avenir and an Eiffel Excellence doctoral fellowship to M. H.

References

Hurtado M et al. (2024). Transcriptomics profiling of the non-small cell lung cancer microenvironment across disease stages reveals dual immune cell-type behaviors. Front. Immunol. 15:1394965. doi: 10.3389/fimmu.2024.1394965

Sturm, G., Finotello, F., Petitprez, F., Zhang, J. D., Baumbach, J., Fridman, W. H., ..., List, M., Aneichyk, T. (2019). Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics, 35(14), i436-i445. https://doi.org/10.1093/bioinformatics/btz363

Benchmarking second-generation methods for cell-type deconvolution of transcriptomic data. Dietrich, Alexander and Merotto, Lorenzo and Pelz, Konstantin and Eder, Bernhard and Zackl, Constantin and Reinisch, Katharina and Edenhofer, Frank and Marini, Federico and Sturm, Gregor and List, Markus and Finotello, Francesca. (2024) https://doi.org/10.1101/2024.06.10.598226

Anna Vathrakokoili Pournara, Zhichao Miao, Ozgur Yilimaz Beker, Nadja Nolte, Alvis Brazma, Irene Papatheodorou, CATD: a reproducible pipeline for selecting cell-type deconvolution methods across tissues, Bioinformatics Advances, Volume 4, Issue 1, 2024, vbae048, https://doi.org/10.1093/bioadv/vbae048

Jiménez-Sánchez, A., Cast, O., & Miller, M. L. (2019). Comprehensive Benchmarking and Integration of Tumor Microenvironment Cell Estimation Methods. Cancer Research, 79(24), 6238–6246. https://doi.org/10.1158/0008-5472.can-18-3560

Morabito S, Reese F, Rahimzadeh N, Miyoshi E, Swarup V. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. Cell Rep Methods. 2023 Jun 12;3(6):100498. doi: 10.1016/j.crmeth.2023.100498. PMID: 37426759; PMCID: PMC10326379.

Langfelder, P., Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008). https://doi.org/10.1186/1471-2105-9-559

Finotello, F., Mayer, C., Plattner, C., Laschober, G., Rieder, D., Hackl, H., ..., Sopper, S. (2019). Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome medicine, 11(1), 34. https://doi.org/10.1186/s13073-019-0638-6

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., … Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nature Methods, 12(5), 453–457. https://doi.org/10.1038/nmeth.3337

Aliee, H., & Theis, F. (2021). AutoGeneS: Automatic gene selection using multi-objective optimization for

RNA-seq deconvolution. https://doi.org/10.1101/2020.02.21.940650

Chu, T., Wang, Z., Pe'er, D. et al. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. Nat Cancer 3, 505–517 (2022). https://doi.org/10.1038/s43018-022-00356-3

Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K. M., Sul, J. H., Pietiläinen, K. H., Pajukanta, P., & Halperin, E. (2020). Publisher Correction: Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nature Communications, 11(1), 2891. https://doi.org/10.1038/s41467-020-16607-9

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M., Melton, D. A., & Yanai, I. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. In Cell Systems (Vol. 3, Issue 4, pp. 346–360.e4). https://doi.org/10.1016/j.cels.2016.08.011

Kang, K., Huang, C., Li, Y. et al. CDSeqR: fast complete deconvolution for gene expression data from bulk tissues. BMC Bioinformatics 22, 262 (2021). https://doi.org/10.1186/s12859-021-04186-5

Frishberg, A., Peshes-Yaloz, N., Cohn, O., Rosentul, D., Steuerman, Y., Valadarsky, L., Yankovitz, G., Mandelboim, M., Iraqi, F. A., Amit, I., Mayo, L., Bacharach, E., & Gat-Viks, I. (2019). Cell composition analysis of bulk genomics using single-cell data. Nature Methods, 16(4), 327–332. https://doi.org/10.1038/s41592-019-0355-5

Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., & Yuan, G.-C. (2019). Accurate estimation of cell-type composition from gene expression data. Nature Communications, 10(1), 2975. https://doi.org/10.1038/s41467-019-10802-z

Xifang Sun, Shiquan Sun, and Sheng Yang. An efficient and flexible method for deconvoluting bulk RNAseq data with single-cell RNAseq data, 2019, DOI: 10.5281/zenodo.3373980

Wang, X., Park, J., Susztak, K., Zhang, N. R., & Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature Communications, 10(1), 380. https://doi.org/10.1038/s41467-018-08023-x

Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C. M., Zou, F., & Jiang, Y. (2020). SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. Briefings in Bioinformatics. https://doi.org/10.1093/bib/bbz166

Zheng SC, Breeze CE, Beck S, Teschendorff AE (2018). "Identification of differentially methylated cell-types in Epigenome-Wide Association Studies." Nature Methods, 15(12), 1059.

Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics. 2013 Apr 15;29(8):1083-5. doi: 10.1093/bioinformatics/btt090. Epub 2013 Feb 21. PMID: 23428642.

Xie, T., Solórzano, J., Madrid-Mencía, M., Essabbar, A., Pernet, J., Kuo, M.-S., Hucteau, A., Coullomb, A., Verstraete, N., Delfour, O., Cruzalegui, F., & Pancaldi, V. (2023). GEM-DeCan : Improved tumor immune microenvironment profiling through novel gene expression and DNA methylation signatures predicts immunotherapy response (p. 2021.04.09.439207). bioRxiv. https://doi.org/10.1101/2021.04.09.439207

Mariathasan, S., Turley, S.J., Nickles, D., Castiglioni, A., Yuen, K., Wang, Y., Kadel, E.E., III, Koeppen, H., Astarita, J.L., Cubas, R., et al. (2018). TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. Nature 554, 544–548.

Gide T.N., Quek C., Menzies A.M., Tasker A.T., Shang P., Holst J., Madore J., Lim S.Y., Velickovic R., Wongchenko M., et al. Distinct Immune Cell Populations Define Response to Anti-PD-1 Monotherapy and Anti-PD-1/Anti-CTLA-4 Combined Therapy. Cancer Cell. 2019;35:238–255. doi: 10.1016/j.ccell.2019.01.003.

Senosain MF et al. (2023). Integrated Multi-omics Analysis of Early Lung Adenocarcinoma Links Tumor Biological Features with Predicted Indolence or Aggressiveness. Cancer Res Commun. 2023 Jul 26;3(7):1350-1365. doi: 10.1158/2767-9764.CRC-22-0373. PMID: 37501683; PMCID: PMC10370362.