nature reviews cancer

https://doi.org/10.1038/s41568-025-00886-9

Supplementary information

A guide to transcriptomic deconvolution in cancer

In the format provided by
the authors and unedited

Supplementary Table 1: Summary of availability scale of public transcriptomic datasets for colorectal, breast, lung, and prostate cancer in GEO (database for bulk and single-cell cohorts), cBioPortal (database for bulk cohorts), and 3CA and TISCH2 (database for single-cell cohorts). The snapshot was taken in December 2024. The summary of all available cancer types in cBioPortal, 3CA and TISCH2 are also provided. Cell lines and patient-derived xenografts are excluded in this summary. Only the curated sets of non-redundant studies were selected in cBioPortal to ensure no overlapping samples. For scRNA-seq studies, only data generated using the 10X platform was selected. Clinical Outcome information availability is determined by the availability of any common survival information, such as overall survival (OS), biochemical relapse-free survival (BCR-free), disease-free survival (DFS) and disease-specific survival (DSS). "-": Information not available. Not shown in the table are the scRNA-seq data availability in clinical trials, based on ClinicalTrials.gov: High-throughput scRNA-seq has been applied to patient samples involved in early clinical trials, although these studies may not explicitly reference scRNA-seq in their trial records. For example, trial NCT02954536, which began in 2016, employed scRNA-seq to study therapy resistance mechanism in 7 metastatic esophageal cancer patients[1]. As of December 2024, over 32 cancer-associated clinical trials have explicitly mentioned single-cell transcriptome or scRNA-seq in their study protocols. However, data availability remains limited: at least 2 trials are marked as completed, but their results have not yet been posted to the database.

	Data Source	Data Type	No. of Studies	No. of Patients	No. of patients with Clinical Outcome (Survival information)
Colorectal Cancer (CRC)	GEO	Bulk approach	>70	>3,400	>1,700
	cBioPortal	Bulk approach	6	1,172	1,162
	GEO	10X scRNA-seq	12	198	64
	3CA	10X scRNA-seq	2	77	-
	TISCH2	10X scRNA-seq	5	36	-
Breast Cancer (BCa)	GEO	Bulk approach	>50	>7,200	>4,900
	cBioPortal	Bulk approach	6	3,582	3,301
	GEO	10X scRNA-seq	24	227	-
	3CA	10X scRNA-seq	7	159	-
	TISCH2	10X scRNA-seq	8	108	-
Lung Cancer (LC)	GEO	Bulk approach	>40	>4,600	>3,200
	cBioPortal	Bulk approach	9	1,615	1,387
	GEO	10X scRNA-seq	22	384	7
	3CA	10X scRNA-seq	8	147	-
	TISCH2	10X scRNA-seq	14	111	-
Prostate Cancer (PCa)	GEO	Bulk approach	>25	>4,300	>3,700
	cBioPortal	Bulk approach	12	1,380	649
	GEO	10X scRNA-seq	10	85	-
	3CA	10X scRNA-seq	2	19	-
	TISCH2	10X scRNA-seq	5	42	-
Multiple Cancer types	cBioPortal	Bulk approach	83	20,574	17,694
	3CA	10X scRNA-seq	76	965	-
	TISCH2	10X scRNA-seq	139	969	-

Supplementary Table 2: Comprehensive overview of computational deconvolution methods mentioned in a recent methods review[2] but not included in our main analysis (**Table 1**). We excluded methods based on four primary considerations: (1) focus on specific non-cancer tissue types limiting applicability to tumor deconvolution; (2) proposed as immediate extension beyond existing approaches; (3) primary application to non-transcriptomic data types (e.g., methylation data); and (4) lack of validation in cancer contexts or over established methods. For each excluded method, we provide the original reference and specific reason for exclusion. This compilation aims to provide transparency in our method selection process while helping researchers understand the broader landscape of available deconvolution approaches beyond those featured in our main analysis.

Category	Method	Reference	Reason for exclusion
Reference based	AdRoit[3]	Yang *et al., Communications Biology* (2021)	Lack of cancer-specific benchmarking.
	ARIC[4]	Zhang *et al. Briefings in Bioinformatics* (2022)	Multi-modal method designed for both transcriptomic and DNA methylation data rather than being specifically focused on transcriptomic deconvolution.
	Bseq-SC[5]	Baron, Maayan *et al., Cell Systems* (2016)	Lack of cancer-specific benchmarking. Validated only on pancreatic islet samples comparing diabetic vs. healthy conditions.
	DCQ[6]	Altboum *et al., Molecular Systems Biology* (2014)	Lack of cancer-specific benchmarking. Validated only on influenza-infected mouse lung time-series (non-cancer context).
	DeconPeaker[7]	Li *et al., Frontiers in Genetics* (2020)	Primarily designed for chromatin accessibility (ATAC-Seq) data rather than transcriptomic data; limited cancer-specific validation.
	DecOT[8]	Liu *et al*., *Frontiers in Genetics* (2022)	Lack of cancer-specific benchmarking (on pancreatic islet pseudo-bulk and real islet bulk RNA-seq data, which are non-cancer tissues).
	DESeq2's unmix[9]	Love *et al., Genome Biology* (2014)	Lack of cancer-specific benchmarking. Repurposes DESeq2's variance-stabilizing transformation.
	DigitalDLSorter[10]	Carlos Torroja and Fatima Sanchez-Cabo. *Frontiers in Genetics* (2019)	Deep learning approach that represents a variation on existing methods without substantial performance improvements in cancer benchmarking compared to established methods.
	dtangle[11]	Hunt *et al., Bioinformatics* (2019)	Lack of cancer-specific benchmarking.
	EMeth[12]	Zhang *et al., Scientific Reports* (2021)	DNA methylation data as input, non-transcriptomic.
	ImmuCellAI[13]	Miao *et al., Advanced Science* (2020)	ssGSEA-based T-cell subset scoring (18 subtypes) without direct regression-based proportion estimates or broader tumor microenvironment benchmarks.
	LinDeconSeq[14]	Li *et al*., *BMC Genomics* (2020)	Variation on an existing approach (represents a minor tweak on DWLS). Limited cancer benchmarking (only applied to AML).
	MethylResolver[15]	Arneson *et al., Communications Biology* (2020)	DNA methylation data as input, non-transcriptomic.

	MySort[16]	Chen et al., *BMC Bioinformatics* (2018)	Validated only on healthy adult PBMC microarray data (benchmark against flow cytometry), with no demonstration on cancer bulk-RNA-seq samples (falls under criterion 1: non-cancer tissue focused).
	spatialDWLS[17]	Dong and Yuan. *Genome Biology* (2020)	Variation on an existing approach (spatial transcriptomics extension of DWLS).
	TICPE[18]	Xiao *et al., Frontiers in Immunology* (2021)	Sample relative expression orderings (REOs) only output absolute cell fraction estimates. Only estimates immune cell types from within cancer tissues, doesn't deconvolve the tumor compartment.
Semi-reference	BisqueMarker[19]	Jew *et al., Nature Communications* (2020)	Semi-supervised model, subset of Bisque. Used when reference profile isn't available.
	DSA[20]	Zhong *et al., BMC Bioinformatics* (2013)	Lack of cancer-specific benchmarking.
	MCP-counter[21]	Becht *et al., Genome Biology* (2016)	GSEA-based method. Only quantifies the abundance of immune and non-immune stromal cell populations. Does not deconvolve tumor compartment.
Reference free	BayCount[22]	Xie *et al., The Annals of Applied Statistics,* (2018)	Deconvolves tumor subclones, doesn't deconvolve tumor proportion or TME.
	BayesCCE[23]	Rahmani *et al., Genome Biology* (2018)	DNA methylation data as input, non-transcriptomic
	CellDistinguisher[24]	Newberg *et al. PLoS ONE* (2018)	Lack of cancer-specific benchmarking.
	debCAM[25]	Chen *et al., Bioinformatics* (2020)	Limited cancer benchmarking. However, it is more advanced and adds probabilistic modeling to account for biological and technical noise.
	DeconICA	Urszula Czerwinska. 2018. R package. No publication	R package. No publication
	ReFACTor[26]	Rahmani *et al. Nature Methods* (2016)	Methylation data as input, non-transcriptomic.
	SMC[27]	Ogundijo *et al., PLoS ONE* (2017)	Lack of cancer-specific benchmarking.

Related links

cbioportal: https://www.cbioportal.org/

Gene Expression Omnibus (GEO): https://www.ncbi.nlm.nih.gov/geo/

Curated cancer cell atlas (3CA): https://www.weizmann.ac.il/sites/3CA/

Tumor immune Single-cell Hub 2 (TISCH2): http://tisch.comp-genomics.org/

References

1. Maron, S. B. *et al.* Determinants of Survival with Combined HER2 and PD-1 Blockade in Metastatic Esophagogastric Cancer. *Clinical Cancer Research* **29**, 3633–3640 (2023).
2. Nguyen, H., Nguyen, H., Tran, D., Draghici, S. & Nguyen, T. Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges. *Nucleic Acids Research* **52**, 4761–4783 (2024).
3. Yang, T. *et al.* AdRoit is an accurate and robust method to infer complex transcriptome composition. *Communications Biology* **4**, 1218 (2021).
4. Zhang, W. *et al.* ARIC: accurate and robust inference of cell type proportions from bulk gene expression or DNA methylation data. *Briefings in Bioinformatics* **23**, bbab362 (2022).
5. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems* **3**, 346-360.e4 (2016).
6. Altboum, Z. *et al.* Digital cell quantification identifies global immune cell dynamics during influenza infection. *Molecular Systems Biology* **10**, 720 (2014).
7. Li, H. *et al.* DeconPeaker, a Deconvolution Model to Identify Cell Types Based on Chromatin Accessibility in ATAC-Seq Data of Mixture Samples. *Frontiers in Genetics.* **11**, (2020).
8. Liu, G., Liu, X. & Ma, L. DecOT: Bulk Deconvolution With Optimal Transport Loss Using a Single-Cell Reference. *Frontiers in Genetics* **13**, (2022).
9. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
10. Torroja, C. & Sanchez-Cabo, F. Digitaldlsorter: Deep-Learning on scRNA-Seq to Deconvolute Gene Expression Data. *Frontiers in Genetics* **10**, (2019).

11. Hunt, G. J., Freytag, S., Bahlo, M. & Gagnon-Bartsch, J. A. dtangle: accurate and robust cell type deconvolution. *Bioinformatics* **35**, 2093–2099 (2019).

12. Zhang, H., Cai, R., Dai, J. & Sun, W. EMeth: An EM algorithm for cell type decomposition based on DNA methylation data. *Scientific Reports* **11**, 5717 (2021).

13. Miao, Y.-R. *et al.* ImmuCellAI: A Unique Method for Comprehensive T-Cell Subsets Abundance Prediction and its Application in Cancer Immunotherapy. *Advanced Science* **7**, 1902880 (2020).

14. Li, H., Sharma, A., Ming, W., Sun, X. & Liu, H. A deconvolution method and its application in analyzing the cellular fractions in acute myeloid leukemia samples. *BMC Genomics* **21**, 652 (2020).

15. Arneson, D., Yang, X. & Wang, K. MethylResolver—a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents. *Communications Biology* **3**, 422 (2020).

16. Chen, S.-H. *et al.* A gene profiling deconvolution approach to estimating immune cell composition from complex tissues. *BMC Bioinformatics* **19**, 154 (2018).

17. Dong, R. & Yuan, G.-C. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biol*ogy **22**, 145 (2021).

18. Xiao, H. *et al.* A Cancer-Specific Qualitative Method for Estimating the Proportion of Tumor-Infiltrating Immune Cells. *Frontiers in Immunology* **12**, (2021).

19. Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications* **11**, 1971 (2020).

20. Zhong, Y., Wan, Y.-W., Pang, K., Chow, L. M. L. & Liu, Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics* **14**, 89 (2013).

21. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology* **17**, 218 (2016).

22. Xie, F., Zhou, M. & Xu, Y. Baycount: A Bayesian Decomposition Method for Inferring Tumor Heterogeneity Using Rna-Seq Counts. *The Annals of Applied Statistics* **12**, 1605–1627 (2018).

23. Rahmani, E. *et al.* BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference. *Genome Biology* **19**, 141 (2018).

24. Newberg, L. A., Chen, X., Kodira, C. D. & Zavodszky, M. I. Computational de novo discovery of distinguishing genes for biological processes and cell types in complex tissues. *PLoS One* **13**, e0193067 (2018).

25. Chen, L. *et al.* debCAM: a bioconductor R package for fully unsupervised deconvolution of complex tissues. *Bioinformatics* **36**, 3927–3929 (2020).

26. Rahmani, E. *et al.* Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nature Methods* **13**, 443–445 (2016).

27. Ogundijo, O. E. & Wang, X. A sequential Monte Carlo approach to gene expression deconvolution. *PLoS One* **12**, e0186167 (2017).