NAR Genomics and Bioinformatics, 2023, **5**, 1–11
https://doi.org/10.1093/nargab/lqad109
Advance access publication date: 23 December 2023
Methods Article

Semi-reference based cell type deconvolution with application to human metastatic cancers

Yingying Lu[1], **Qin M. Chen**[2,3] **and Lingling An**[1,4,5,*]

[1]Interdisciplinary Program in Statistics and Data Science, University of Arizona, Tucson, AZ, USA
[2]College of Pharmacy, University of Arizona, Tucson, AZ, USA
[3]Cancer Biology Program, University of Arizona, Tucson, AZ, USA
[4]Department of Biosystems Engineering, University of Arizona, Tucson, AZ, USA
[5]Department of Epidemiology and Biostatistics, University of Arizona, Tucson, AZ, USA

[*]To whom correspondence should be addressed. Tel: +1 520 621 1248; Email: anling@arizona.edu

Abstract

Bulk RNA-seq experiments, commonly used to discern gene expression changes across conditions, often neglect critical cell type-specific information due to their focus on average transcript abundance. Recognizing cell type contribution is crucial to understanding phenotype and disease variations. The advent of single-cell RNA sequencing has allowed detailed examination of cellular heterogeneity; however, the cost and analytic caveat prohibits such sequencing for a large number of samples. We introduce a novel deconvolution approach, SECRET, that employs cell type-specific gene expression profiles from single-cell RNA-seq to accurately estimate cell type proportions from bulk RNA-seq data. Notably, SECRET can adapt to scenarios where the cell type present in the bulk data is unrepresented in the reference, thereby offering increased flexibility in reference selection. SECRET has demonstrated superior accuracy compared to existing methods using synthetic data and has identified unknown tissue-specific cell types in real human metastatic cancers. Its versatility makes it broadly applicable across various human cancer studies.

Introduction

Bulk RNA-seq experiments are routinely employed to characterize gene expression patterns by aggregating the information from a tissue or organ containing various cell types. Standard methods such as differentially expressed gene analysis or eQTL studies rely on a population level relationship that leverages an overall gene abundance for each sample but neglects cell type specific information (1). Notably, cell type compositions and proportions for each sample exhibit considerable diversity. Comprehending the existence of such diversity and interactions between cell types is vital since these may influence variations in gene expression. For instance, individuals diagnosed with the same disease might exhibit different phenotypes or drug responses, a heterogeneity potentially linked to variations in cell type proportion or cell type-specific gene expression levels (2). Investigating cell type heterogeneity is essential for deeper understanding of tissue characteristics, pinpointing biological alterations or treatment options in disease and dysfunction, such as in cancers (2–4).

Traditional technologies such as flow cytometry or immunohistochemistry (IHC) are standard methods to capture cell type specific information (5). However, these techniques require specialized costly equipment, are labor intensive, and most importantly can only measure a limited number of characteristics concurrently in a defined two dimensional space. These methods lack scalability and adaptability compared to the current sequencing based data generation (6). In recent years, single-cell RNA sequencing technologies have emerged, becoming increasingly popular for addressing biological questions. Despite their potential to derive cell type information, these methods have their own biases and limitations (7). Primarily, they necessitate highly trained laboratory technicians and individual cell separation technologies, thereby introducing potential human errors. Second, the procedures and equipment are expensive and time-consuming, rendering this method unsuitable for large-scale clinical applications. Furthermore, certain types of tissues are not suitable for single cell sequencing, such as solid tissue (8), or tightly packed stromal cells. In contrast, the bulk RNA-seq studies over the past decade have generated an abundance of datasets; and these invaluable resources can be utilized to investigate various medical conditions, drug effects, or time-series studies. Under most circumstances, the biological tissues or samples used in RNA-seq are not duplicated for laboratory experimental testing (9). The relatively low cost and high throughput capacity of bulk RNA-seq exceed that of single cell RNA-seq. As a result, the population based RNA-seq data are increasingly abundant, supporting the importance of developing advanced methods for determining cell type composition.

Cell type proportion estimation, often referred to as cell type deconvolution, can generally be divided into two categories based on their input data: reference-based and reference-free (10). Reference-free methods do not require actual cell type identification as a reference, and instead directly decompose the bulk data. This approach, suffers from a high error rate and can lead to ambiguity in subsequent cell type annotation (6). In contrast, reference-based methods utilize information from predefined cell type expression datasets

Received: July 27, 2023. **Revised:** November 1, 2023. **Editorial Decision:** December 11, 2023. **Accepted:** December 13, 2023

or well-established marker genes, therefore are heavily reliant on the accuracy of the reference dataset. Comparative studies evaluating these two methods have revealed that using a reference can enhance accuracy and provide meaningful cell type explanations (6,11). However, sometimes, not all cell type information in the reference may be available or precise, especially when predicting cell types in cancer samples due to the ununiform nature. Utilizing healthy donors or even tumor samples as references could overlook information about specific cancer cells or miss certain cell types due to progression stages (12). Further complications include that primary tumor sites host cells with diverse genetic, epigenetic, and phenotypic characteristics, resulting in a mixed population with different molecular characteristics and vulnerability to treatment (13). When these cells metastasize, they carry this heterogeneity and may acquire new traits that further diversify the cell population in a metastatic tumor, allowing for adaptation to local tissue environments. Therefore, a metastatic tumor may contain a cell type not found in the primary tumor or the reference dataset, pointing to the limitation of current reference-based methods in comprehensively capturing tumor cell heterogeneity.

Currently, most reference based cell type deconvolution methods struggle to accommodate a situation where a cell type is present in the tissue for generating the bulk data but is absent in the reference (10). Two published methods on cell type deconvolution, EPIC (14) and PREDE (10), are capable of addressing such scenarios involving missing cell types. EPIC encompasses a reference curated from RNA-seq-based gene expression derived from major immune and other non-malignant cell type. This enables the prediction of both cancer and immune or other non-malignant cell types from bulk gene expression data. PREDE, in contrast, can infer proportions of multiple unknown cell types by solving a Non-negative Matrix Factorization (NMF) model. However, both EPIC and PREDE lack references built from single cell sequencing data, which now provide detailed and reliable information about cellular heterogeneity and distinct cell type characterization. Additionally, these two methods utilize the square error (L2 loss) function, which is sensitive to outliers and less robust compared to the absolute error (L1 loss).

We propose a novel framework employing SEmi-referenCe generated from single-cell RNA-seq data to Estimate cell Type proportions (SECRET). This semi-reference-based method leverages partial reference containing gene expression to estimate shared cell types between bulk data and reference, while also allowing the estimation of unknown cell type(s) absent in the reference. Our method is benchmarked using synthetic bulk samples simulated from real single-cell data, where true cell type proportions are predefined, and further applied to various metastatic cancers.

Materials and methods

Overview of SECRET

A schematic overview of SECRET is presented in Figure 1. SECRET is built on the commonly used formulation of cell type deconvolution: $Y = CP$, where Y represents expression matrix of M genes in N samples (2,11). C is the cell type expression treated as reference data for cell type estimation. C consists of M genes by K cell types. P is the proportions of K cell types for N samples. P is our target to be estimated. Sometimes the C matrix is not complete due to biological differences between the bulk samples and the reference samples. For example, if primary tumor samples are used to decompose the metastatic cancer samples, even with the matched patients, the cell types in the metastatic site are expected to be somehow different from the primary site due to cell type evolvement by metastasis (15). In practical applications, it's essential to consider the potential presence of unknown cell types, which may be specifically associated with certain tissues, when conducting cell type deconvolution analyses.

Cell type proportion estimation

Assume there are K cell types profiled from reference data, and for generality, we assume there is one unknown cell type contained in bulk tissue but not included in the reference data. The primary process of SECRET seeks to find the optimal proportion estimation for $K + 1$ cell types. The objective is to minimize the discrepancy between the estimated bulk gene expression and the observed bulk gene expression (10,11,16). This can be solved through a weighted constrained nonlinear optimization for a bulk RNA-seq sample:

$$\min_{P}\left\{\sum_{i=1}^{M} w_i * \left| y_i - \sum_{k=1}^{K+1} p_k c_{ik} \right|\right\}$$

where y_i is the gene i expression, $\boldsymbol{p}$ corresponds to the proportions, comprising of $K + 1$ cell types. These include K cell types identified from single cell reference and an additional unknown cell type found in the bulk sample but absent in the reference. Particularly, for each bulk sample, $p_k \geq 0,\ k = 1,\ \ldots,\ K+1$, and $\sum_{1}^{K+1} p_k = 1$. When $k = 1,\ \ldots,\ K$, the term c_{ik} denotes the expression for gene i in cell type constructed from single-cell data. When $k = K + 1$, the c_{ik} represents the gene expression profile for the unknown cell type that needs to be estimated.

SECRET adopts absolute deviation loss due to its resilience against outliers. The optimization involves not just minimizing the objective function, it's also constrained by dual conditions. These conditions ensure that each proportion is nonnegative and that the total across cell types sums up to one within each sample. The optimization phase utilizes the augmented Lagrangian method, good for its efficacy in handling nonlinear inequality constraints. In the context of an unknown cell type, our algorithm initiates the process by assigning a zero expression profile to accommodate the initial absence of data. This initiation is followed by a complexed, iterative optimization stage, recursively fine-tuning these preliminary values. Throughout this stage, the model is designed to minimize discrepancies between the observed bulk gene expressions and the estimations based on single-cell data. Further elaboration on the reference construction steps is available in the supplementary file. Central to this stage is the enforcement of two stringent constraints: each cell type proportion is nonnegative, and the sum of proportions of all cell types, known and unknown, must equal one. The iterations persist until the algorithm converges on a solution within a predefined tolerance level. By balancing mathematical rigor with biological fidelity, our approach presents a reliable framework for decomposing cellular compositions, even in scenarios complicated by unknown elements.

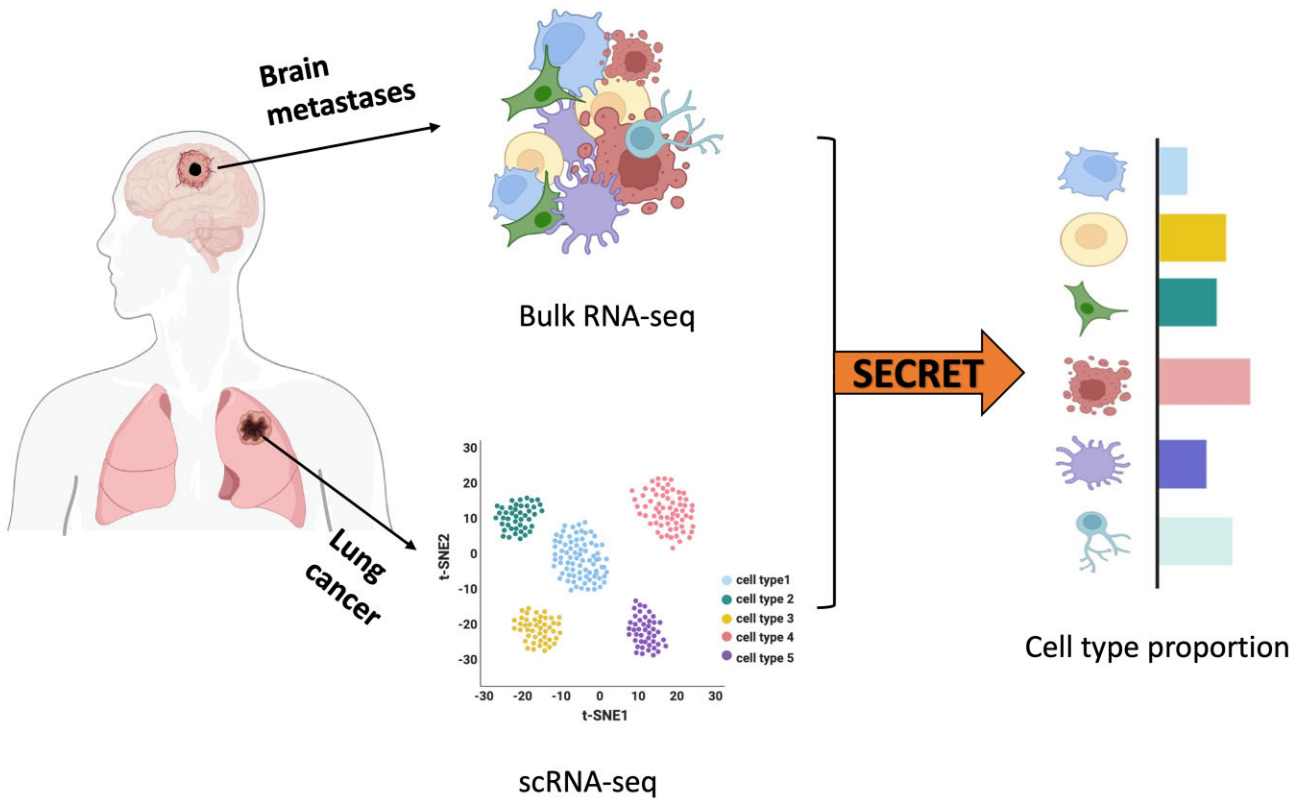

Figure 1. Schematic overview of SECRET. SECRET uses both bulk RNA-seq and scRNA-seq as input, by solving a constraint nonlinear optimization problem to find a set of cell type proportions that minimize the relative error between true gene expression and estimated gene expression.

Weight assignment

To differentiate the contribution of each gene, we design to assign weight denoted by w_i to the gene i $(= 1, \ldots, M)$. The selection of M genes based on such criteria. Any genes shared between single-cell data and bulk data that fulfill either of the subsequent conditions will be retained: (a) commonly recognized cell type markers or frequently referenced markers from prior publications, and (b) differentially expressed genes (DEGs). In our study, the identification of differentially expressed genes (DEGs) was an important component, achieved by comparing each cell type to all others within the single-cell data. This approach, common in the field for its efficacy in discerning markers that define cell types, was implemented through the ‘FindAllMarkers’ function inherent to the Seurat package (17). The detect of DEG involved conducting a likelihood-ratio test (16), focusing on genes observable in a certain percentage of cells across the cell types under comparison. We imposed a restriction in our analysis to consider only those genes that were present in a minimum of 25% of cells in any of the two comparative groups, ensuring the statistical significance and relevance of the genes analyzed. Beyond these specifications, we adhered to the default parameters within the function, thereby aligning our methodology with standard practices and enhancing the reproducibility of our study (18). Notably, while some DEGs were cell type-specific, others served as more general markers but were no less integral to our analysis.

We calculate weights for gene i as $w_i = 1/F_i$, where F_i represents the frequency of gene i occurrence between cell types. We identify a set of markers for each cell type, following the gene selection procedures outlined in the previous section. These markers are not necessarily unique to a single cell type. If two cell types are closely related, they may share similar markers. Consequently, overlapping genes may not serve as effective discriminators between these two cell types. The underlying rationale for this weight assignment is that a gene differentially expressed in a unique cell type should carry more information for distinguishing that cell type from the others, compared to genes differentially expressed among multiple cell types. Therefore, the weight we assign to a gene is inversely related to its frequency of occurrence across different cell types. For instance, if gene A is a marker for three cell types, also referred to as a non-specific gene, and gene B, a specific gene, which is exclusively expressed in a cell type, gene A would be assigned a less weight than gene B. The weights for gene A and gene B, therefore, would be 1/3 and 1, respectively. This weighted scheme enables a quantifiable assessment of the unique contribution of each gene to a cell type.

Simulation studies

In order to evaluate the efficacy of our proposed approach, we conducted simulations using pseudo-bulk samples that were derived from actual single-cell RNA-seq data of human metastatic lung adenocarcinoma (19). Lung cancer, characterized by uncontrolled cell growth in lung tissues, is among the most common and deadly types of cancer worldwide. Its severity is exacerbated by its propensity to metastasize or spread to other parts of the body, with the brain being one of the most common sites of lung cancer metastasis. Brain metastases occur when cancer cells detach from the primary lung tumor and migrate via the circulation system to the brain, where they can proliferate and form new, secondary tumors possessing different characteristics from the primary lung tumor. The brain consists of unique cell types, such as oligodendrocytes, which are absent in the lung (19). Hence, our simulation employed brain metastases to construct pseudo-bulk data by summarizing gene expression across all cell types and subjects.

The generation of bulk data involved random selection of eight brain tumors from a total of ten samples. Therefore, when we repeat the simulation, we are able to construct a variety of sample datasets to validate our algorithm. For each tumor sample, 70% of the cells from each of the eight cell types, myeloid cells, T/NK cells, B lymphocytes, fibroblasts, endothelial cells, MAST cells, epithelial cells, oligodendrocytes, were used. By accounting for the number of cells employed for each tumor, we could ascertain cell type proportions as the ground truth. Subsequently, we aggregated gene counts from the chosen cells across all cell types to attain mixed gene expression for each tumor sample. The single-cell RNA-seq data used to construct the cell type reference was derived from the same study that consist of 11 primary tumor samples and 38489 cells. The primary lung tumors contained seven cell types, identical to those from the brain site, except the oligodendrocytes - a cell type unique to the central nervous system, inclusive of the brain and spinal cord (20).

We conducted our evaluation by comparing our new method to EPIC and PREDE across three different simulation settings. In the first setting, we randomly generated a pseudo-bulk and use SECRET to estimate its cell type composition. In the second setting, to account for inherent randomness in bulk data generation, we repeated the simulation ten times using different random seeds, with no additional noise in this scenario. Thus, each replicate may consist of a different tumor samples and different cells from each cell type. In the third setting, we introduced noise into the synthetic bulk data. This noise was generated based on a normal distribution $N(0,\ a * Y)$ where a takes values $c(0,\ 0.05,\ 0.1,\ 0.2,\ 0.4)$ and Y is the synthetic bulk data. To evaluate the performance, we used the most commonly cited metrics: mean absolute deviation (mAD), $\text{avg}(|\boldsymbol{p} - \hat{\boldsymbol{p}}|)$, root mean squared deviation (RMSD), $\sqrt{\text{avg}(\boldsymbol{p} - \hat{\boldsymbol{p}})^2}$, and Pearson correlation, $\text{Cor}(\boldsymbol{p},\ \hat{\boldsymbol{p}})$, where $\boldsymbol{p}$ is the vector of the true cell type proportions and $\hat{\boldsymbol{p}}$ is the vector of the estimated cell type proportions.

Results

Performance on simulated lung adenocarcinomas

For setting 1, we utilized one replicate of SECRET estimation to compare the estimated cell type proportions with their true proportions, as illustrated in a scatter plot (Figure 2). The dots align well along the 45-degree line, suggesting a strong correlation between the estimated and true proportions. Upon examining the cell type within each sample, we represented the results as a heatmap (Supplementary Figure S1). SECRET exhibits a pattern that closely matches the true proportions, especially excelling in detecting dominant cell types.

For setting 2, we regenerated the pseudo-bulk data ten times without adding any extra noise to evaluate the robustness of SECRET. Each new bulk data set composed of varied samples and cell sets from each cell type. As illustrated in Figure 3A, each boxplot represents the results of ten replicates per method, with SECRET consistently outperforming EPIC and PREDE by displaying the lowest deviance and highest correlation.

For setting 3, an additional four levels of noise were applied to the pseudo-bulk data for one replicate to further assess the performance of SECRET. Figure 3B shows that as the noise levels increase, the overall performance of both SECRET and EPIC experiences a minor decline. Across all noise levels, SECRET consistently exhibits the lowest mAD and RMSD, while maintaining the highest correlation coefficient (R).

Application to breast cancer brain metastases

Breast cancer is a leading disease in women, categorized into three subtypes: (i) Luminal (ER+), (ii) HER2+ and (iii) triple-negative, depending on the status of specific receptors (TNBC) (21). The aggressive metastatic form, breast cancer brain metastases (BCBM), illustrates the adaptability of breast cancer cells to different therapeutic pressures and environments, influenced by estrogen receptor (ER) status (22). BCBMs show considerable genomic and phenotypic differences from their primary tumors, with the genomic variance particularly prominent across tumor subtypes (23).

To assess the variance in cellular composition between primary breast tumors and brain metastases, we leveraged a dataset derived from transcriptomics of both primary breast tumors and brain metastases, from each subtype of breast cancer (23). For each of the cancer subtypes, ER+, HER2 + and TNBC, we built cell type reference using single-cell RNA-seq data from human primary breast tumors, comprising 11, 5 and 10 samples respectively (21). Using SECRET to decompose bulk data of both the primary breast tumor and brain metastases, Figure 4 shows a consistent decrease in 6 out of 10 cell types across all subtypes, particularly CAFs cancer associated fibroblast cells (CAFs), which significantly impact tumor progression from primary breast tumors. These cells promote tumor growth, migration, and invasion through extracellular matrix remodeling, growth factor secretion, and immune response modulation (24). Changes in the tumor microenvironment when breast cancer metastasizes to the brain could result in a lower proportion of CAFs (25). The brain's microenvironment includes brain cells like neurons, astrocytes, oligodendrocytes, and microglia, which are absent in primary breast tumors and may affect tumor growth and progression in the brain (26). The role of these brain cells in metastatic tumors is a valuable avenue for understanding the interaction of breast cancer cells with local environment.

The subtypes of ER+, HER2+ and TNBC have distinct immune cell proportions. ER+ has fewer B-cells, T-cells, and plasmablasts than HER2+ and TNBC, as reported (27–29). HER2+ and TNBCs are more immunogenic, attributed to their higher mutation rates and increased production of neoantigens, drawing more immune cells (30). ER+ breast cancers, less immunogenic with fewer immune cells, are influenced by an active hormone receptor pathway, causing immunosuppression and diminished immune infiltration (31,32).

Application to colorectal cancer liver metastases

About 25–50% of colorectal cancer (CRC) patients develop metastases, often detected during the diagnosis, primarily in the liver (33,34). Utilizing bulk samples from a pilot study (34), we examined the cell types in metastatic tumors of two CRC patients with liver metastases. The first patient unfortunately passed away from tumor growth despite treatment,

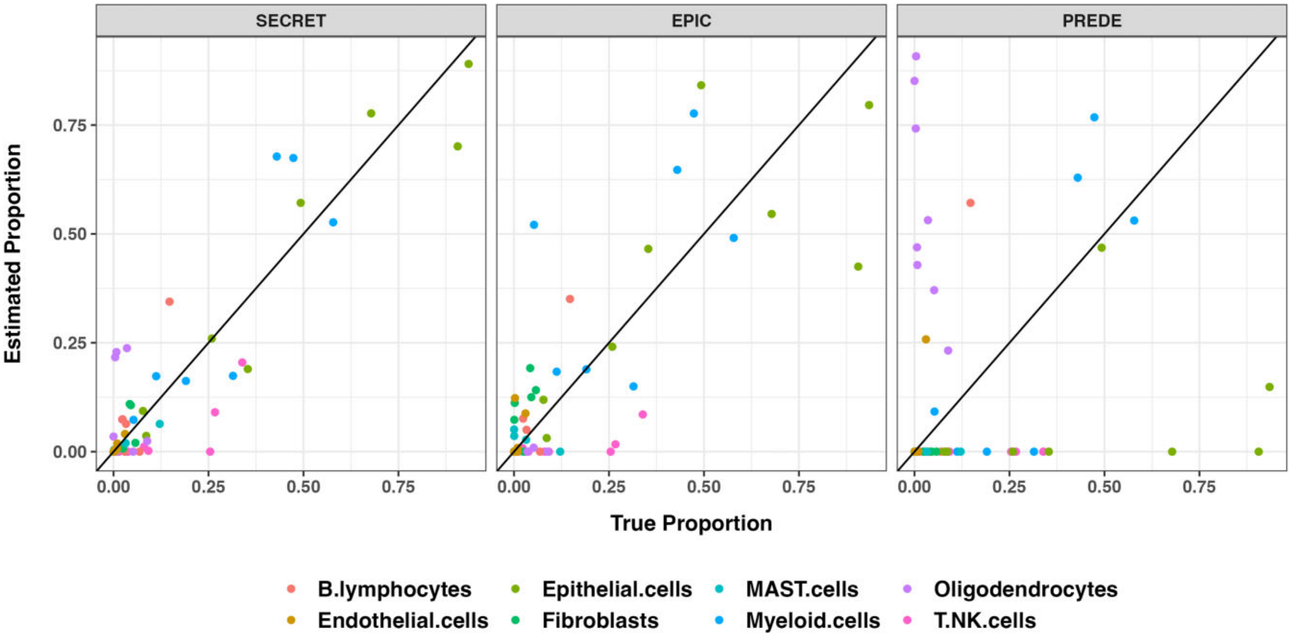

Figure 2. Comparison between estimated and true cell type proportion from simulation study. Each colored dot represents a cell type for a sample, 45 degree lines indicated the same proportions between estimated and true proportion.

while the second survived. The study employed a single-cell analysis of primary treatment-naive colorectal cancer. After data filtering, it incorporated 26 samples and identified nine main cell types (35).

After SECRET estimation, we observed varying immune cell responses between two distinct patients (Figure 5A). Patient 1 displayed reduced immune cells, encompassing B cells, T cells, innate lymphoid cells (ILCs), plasma cells, and myeloid cells, all fundamental to immune responses (33,36,37). The reduced immune cell presence in patient 1 could be attributed to immunosuppression, a tumor-induced condition that impairs immune cell function and growth (38). Alternatively, immune cell exhaustion could be another factor, diminishing the effectiveness of cells like T cells against tumors (39). The different outcomes between these two patients could be partly due to patient 1′s lowered immune cell levels, possibly accelerating cancer progression.

To investigate whether the unidentified cell type is related to liver tissue-specific cell types, specifically hepatocytes, we examined single-cell RNA-seq data from five healthy human liver samples (36). Hepatocytes are the primary functional cells in the liver, responsible for most of the organ's metabolic functions (37). The acquired data comprises 8848 cells, spanning 14 different cell types. We utilized the same constrained nonlinear optimization framework of SECRET to evaluate the cell type-specific gene expression, based on the estimated proportions derived from SECRET and the bulk data. Upon mapping the gene expression of this unidentified cell type for patient 1 onto the UMAP generated from liver samples (Figure 5B), the gene expression corresponding to the unknown cell type fell within the hepatocyte range. This finding substantiated our hypothesis that this unidentified cell type is related to the liver. Additionally, a heatmap was produced using well-established hepatocyte marker genes (Figure 5C), which revealed the hepatocyte marker genes are highly expressed in the unidentified cell type for patient 1. Similar results were obtained for patient 2, as detailed in Supplementary Figure S2 and S3.

Application to pancreatic ductal adenocarcinoma (PDAC)

Pancreatic ductal adenocarcinoma (PDAC), making up about 90% of all pancreatic cancers, is notorious for its early metastasis and high mortality rate, particularly due to liver metastasis (40,41). PDAC liver metastases are often immune to standard chemotherapy and radiation therapy, due to PDAC's aggressiveness, complex tumor microenvironment, and highly treatment-resistant nature (42,43). Our aim was to utilize the information derived from the primary cell types to forecast the cell type composition in hepatic metastases and to draw a comparison between primary and metastatic tumors. Information regarding cell types was built from the primary tissue (44), comprising 24 samples and 9 distinct cell types. We downloaded the bulk data of 13 primary and 14 metastatic tumors from a previous study (45). Through the implementation of cell type deconvolution using the SECRET algorithm, we noted a contrasting pattern of change between acinar cells and stellate cells (Figure 6A). Pancreatic acinar cells, unique to the pancreas, were unsurprisingly not estimated at metastatic sites (46), whereas stellate cells, pivotal in the development of liver fibrosis were present (47). A dominant, unidentified cell type was estimated in liver metastasis, exhibiting a strong correlation with stellate cells and B cells (Figure 6B). Interactions were found with certain liver-related cell types, such as hepatocytes, which are known to interact with stellate cells, particularly in the context of liver injury or disease (48). To test our hypothesis, we mapped this unidentified cell type onto the scRNA-seq derived from healthy liver samples (36). The subsequent UMAP visualization (Supplementary Figure S4) suggested a relationship between this unidentified cell type and hepatocytes.

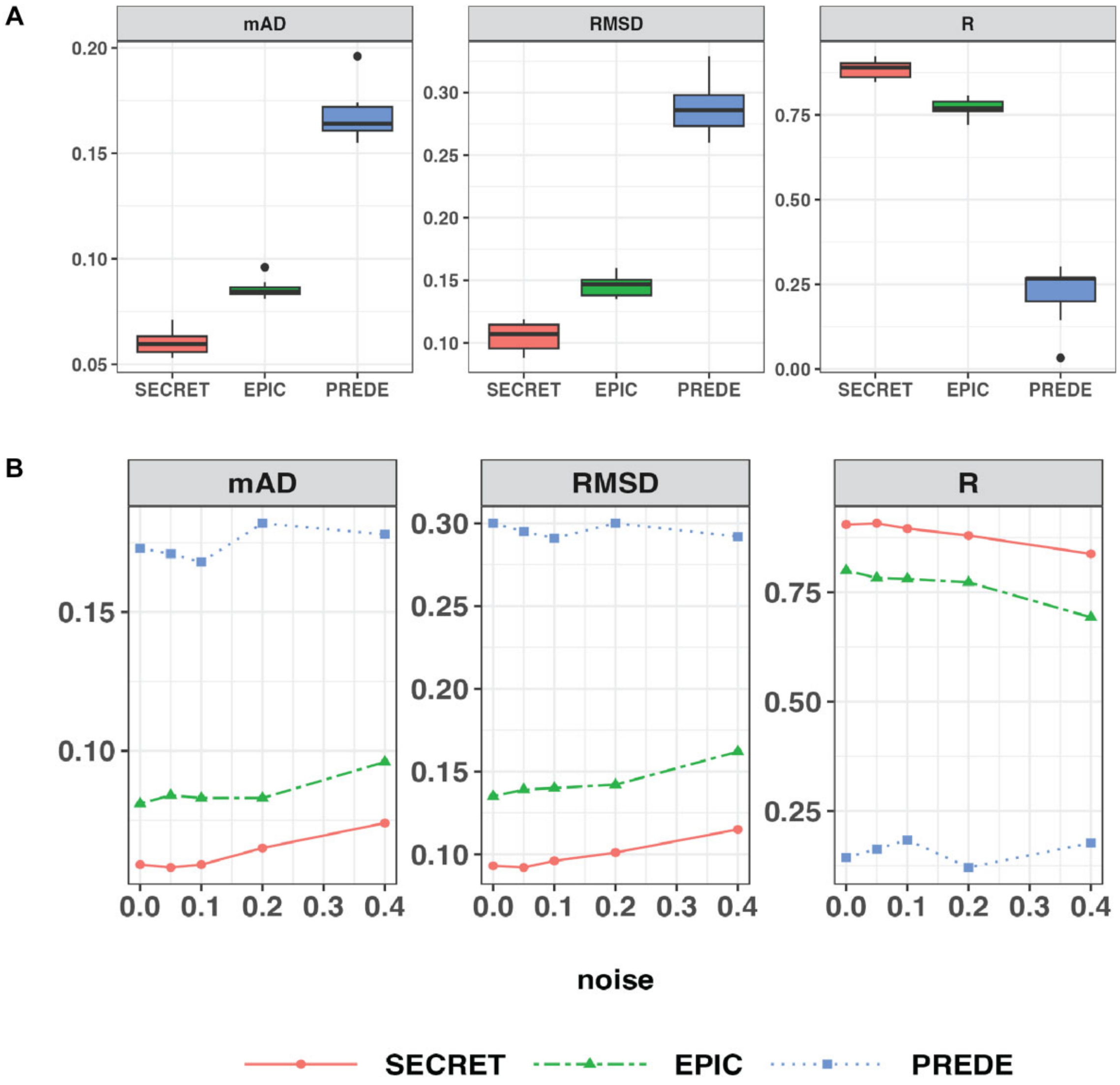

Figure 3. Simulation performance. (**A**) Boxplot comparing the performances of SECRET, EPIC and PREDE when one cell type is unidentified. (**B**) Assessment for various noise levels. The x-axis represents noise levels, and the y-axis signifies the values corresponding to each evaluation metric. Each panel illustrates a distinct evaluation outcome. Results obtained from SECRET, EPIC and PREDE are represented by colored lines in red, green, and blue respectively. Lower values are desirable for metrics like mAD and RMSD, while a higher value is preferable for the correlation coefficient, denoted as R.

We analyzed the high-expressed genes in the unknown cell type with Gene Ontology for biological process, cellular component, and molecular function (Figure 6C). In pancreatic ductal adenocarcinoma (PDAC) liver metastases, Gene Ontology shows increased inflammatory responses, metabolic shifts, and enzyme regulation, all contributing to cancer cell survival and growth (49). Cellular components like blood microparticles, collagen-rich extracellular matrix, and high-density lipoprotein particles indicate environmental changes tied to metastasis (50). Altered binding activity and peptidase function suggest proteolytic changes involved in metastasis (51). These findings highlight the interaction of inflammation, metabolic adaptation, and matrix remodeling in PDAC liver metastases.

Discussion

Cellular deconvolution of bulk RNA-seq data represents a research domain with intense interest. In this field, researchers are exploring new methods, including deep learning algorithms (52,53) and Bayesian modeling (54), to better understand cell compositions from RNA-seq data. These methods are becoming popular because they can analyze complex biological data more effectively. For instance, deep learning uses artificial neural networks to study large datasets and find patterns related to different cell types. Likewise, Bayesian methods utilize probability to estimate cell compositions, incorporating uncertainties and leveraging existing knowledge to enhance the precision of their findings. Furthermore, we conducted an additional comparison, including a Bayesian method called BayICE (54), and the results demonstrate that our proposed method continues to outperform all other approaches (Supplementary Figure S5). Yet, relatively few algorithms have been developed that can effectively manage cell types not present in the reference dataset. Although reference-free methodologies may have the potential to tackle unknown cell types, they encounter with challenges related to cell type identification in the absence of prior knowledge. Despite these

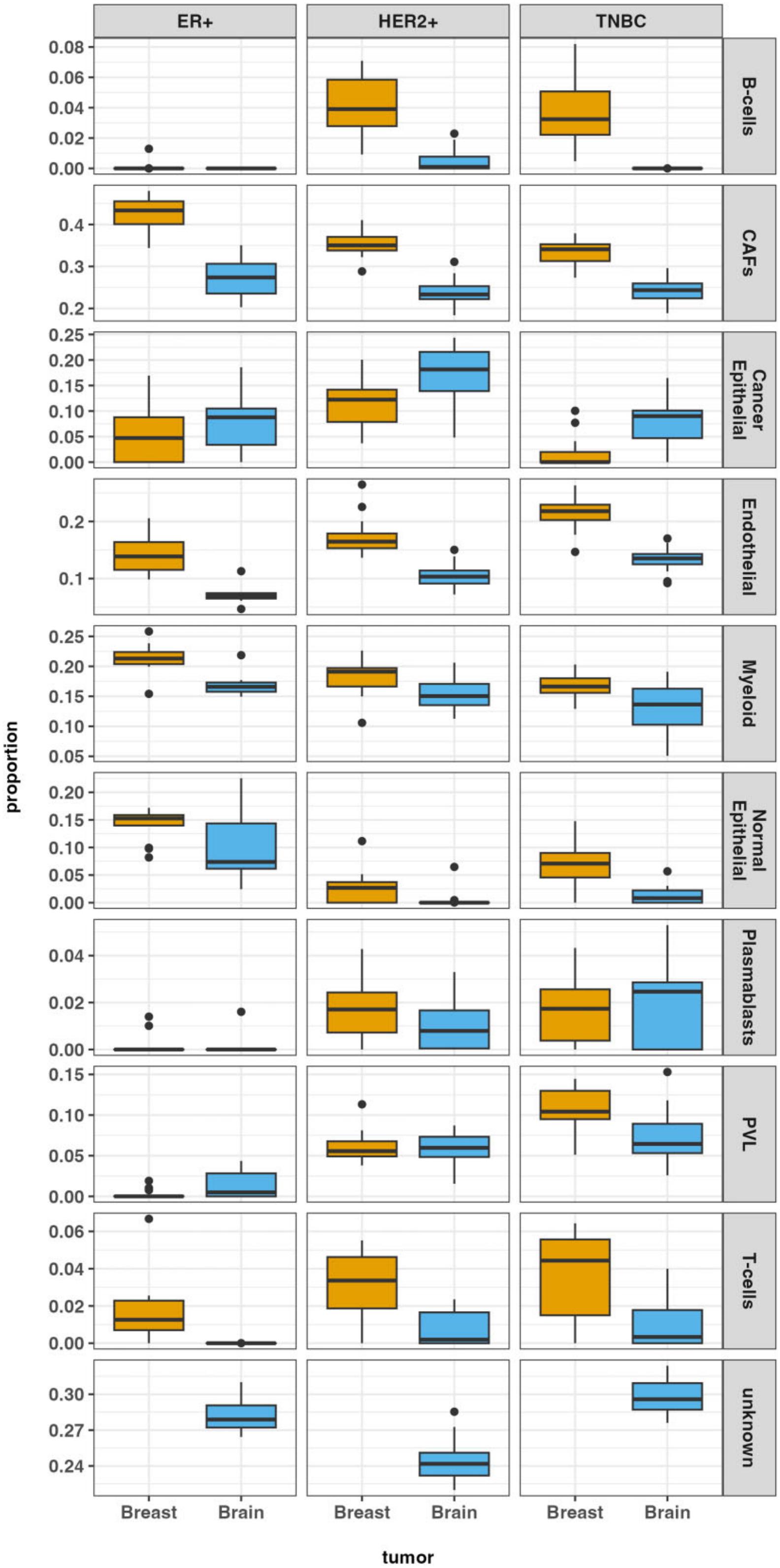

Figure 4. Cell type proportion inferred by SECRET for breast tumors and brain metastases within each breast cancer subtype.

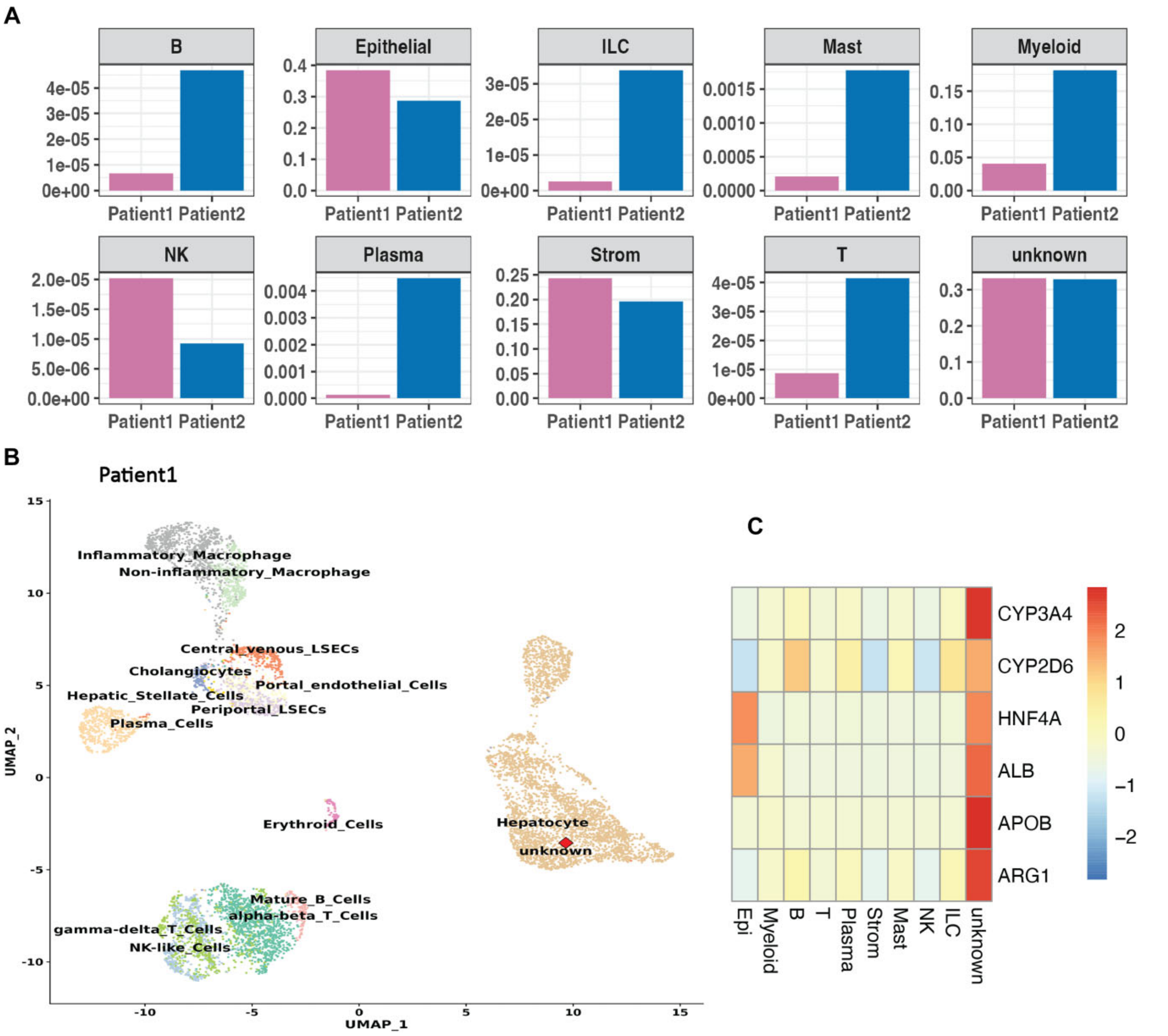

Figure 5. Cell type estimation results for colorectal cancer liver metastases. (**A**) Estimated proportion for each cell types for two patients. (**B**) An unknown cell type for patient1 is plotted against the single-cell RNA-seq data from five liver samples. The positioning suggests a strong relation between the unknown cell type with hepatocytes. (**C**) Heatmap of scaled estimated gene expression for hepatocyte markers in both the referenced cell types and the unknown type.

limitations, reference-based methodologies continue to be a favored approach. Our newly proposed algorithm, SECRET, optimally fuses the advantages of both reference-based and reference-free methodologies. This method employs single-cell RNA-seq data to guide the cellular deconvolution process, providing identifications for expected cell types. Importantly, SECRET also makes allowances for unknown cell types not represented in the reference data, achieved by adjusting algorithmic parameters.

Simulation studies have underscored the superior performance of SECRET, particularly in scenarios with the presence of unknown cell types. In the context of various metastatic cancer datasets, single-cell studies have indicated significant variations in the cellular composition between primary tumor sites and metastatic locations. Moreover, SECRET's scalability and efficiency are assured. In both simulations and real data applications, we utilized a range of 8 to 10 cell types, demonstrating its capability to scale to a larger number of cell types. In contrast, EPIC and PREDE limited their methods to datasets containing 3–5 cell types. Beyond its scalability, SECRET stands out for its operational efficiency. Our comparison with well-known methods like EPIC and PREDE, detailed in our supplementary table, shows that SECRET performs better. It not only gives the most accurate results, aligning closely with the actual proportions, but it also does this faster than the others, having the shortest processing time. This combination of precision and expedited data processing establishes SECRET as an invaluable tool for applications that require rapid and reliable insights. By utilizing SECRET, we successfully identified an unknown cell type, correlating it with tissue-specific cells, underscoring the algorithm's vast potential for a wide range of biological and clinical applications.

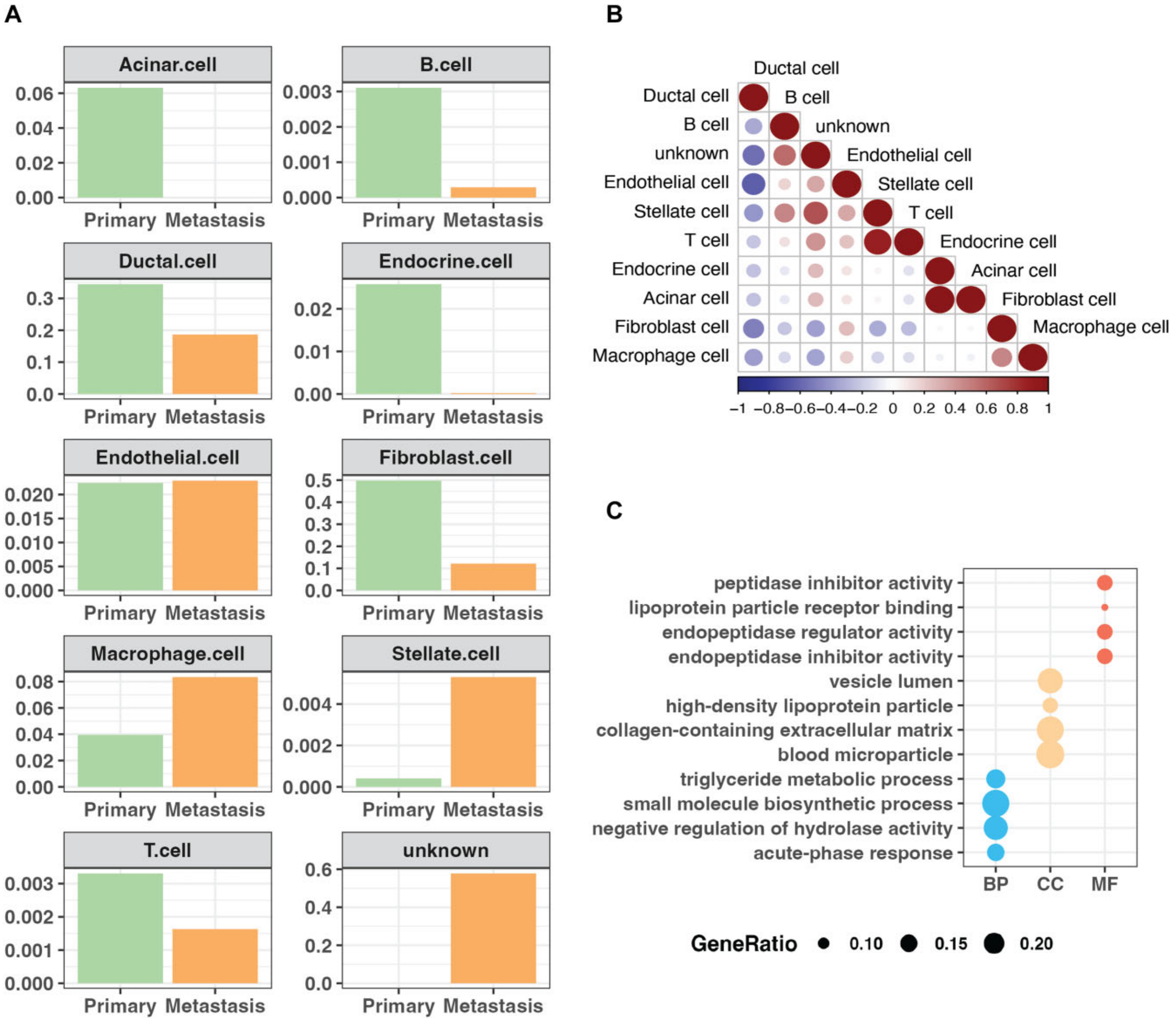

Figure 6. Results for pancreatic ductal adenocarcinoma (PDAC) liver metastases. (**A**) Estimated cell type proportions of both primary and metastatic tumors from SECRET. (**B**) Correlations between cell type proportions for liver metastases. (**C**) Top GO terms based on the highly expressed genes from unknown cell type.

Data availability

All datasets used in this study are publicly available. The single cell RNA-seq of lung adenocarcinoma can be accessed using code GSE131907. The bulk and scRNA-seq of BCBM have access code GSE184869 and GSE176078. The bulk and scRNA-seq of Colorectal Cancer Liver Metastases have access code GSE162960 and GSE178341. The bulk data of PDAC have accession number GSE151580 and the corresponding scRNA-seq can be found in the Genome Sequence Archive under project PRJCA001063. An R package has been developed to implement SECRET available from Zenodo https://doi.org/10.5281/zenodo.8157419 and GitHub at https://github.com/anlingUA/SECRET.

Supplementary Data

Supplementary Data are available at NARGAB Online.

Funding

National Institutes of Health [R01 GM125212, GM126165]; Holsclaw endowment to Q.M.C; National Institutes of Health [R01 GM139829, P01 AI148104-01A1]; United States Department of Agriculture [ARZT-1361620-H22-149] to L.A.

Conflict of interest statement

None declared.

References

1. Jew,B., Alvarez,M., Rahmani,E., Miao,Z., Ko,A., Garske,K.M., Sul,J.H., Pietiläinen,K.H., Pajukanta,P. and Halperin,E. (2020) Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat. Commun.*, **11**, 1971.

2. Wang,X., Park,J., Susztak,K., Zhang,N.R. and Li,M. (2019) Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.*, **10**, 380.
3. Dong,M., Thennavan,A., Urrutia,E., Li,Y., Perou,C.M., Zou,F. and Jiang,Y. (2021) SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief. Bioinform.*, **22**, 416–427.
4. Tsoucas,D., Dong,R., Chen,H., Zhu,Q., Guo,G. and Yuan,G.C. (2019) Accurate estimation of cell-type composition from gene expression data. *Nat. Commun.*, **10**, 2975.
5. Sturm,G., Finotello,F., Petitprez,F., Zhang,J.D., Baumbach,J., Fridman,W.H., List,M. and Aneichyk,T. (2019) Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics*, **35**, i436–i445.
6. Jin,H. and Liu,Z. (2021) A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biol.*, **22**, 102.
7. Lähnemann,D., Köster,J., Szczurek,E., McCarthy,D.J., Hicks,S.C., Robinson,M.D., Vallejos,C.A., Campbell,K.R., Beerenwinkel,N., Mahfouz,A., *et al.* (2020) Eleven grand challenges in single-cell data science. *Genome Biol.*, **21**, 31.
8. Gao,M., Guo,P., Liu,X., Zhang,P., He,Z., Wen,L., Liu,S., Zhou,Z. and Zhu,W. (2022) Systematic study of single-cell isolation from musculoskeletal tissues for single-sell sequencing. *BMC Mol. Cell Biol.*, **23**, 32.
9. Hicks,S.C., Townes,F.W., Teng,M. and Irizarry,R.A. (2018) Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostat. Oxf. Engl.*, **19**, 562–578.
10. Qin,Y., Zhang,W., Sun,X., Nan,S., Wei,N., Wu,H.J. and Zheng,X.. (2020) Deconvolution of heterogeneous tumor samples using partial reference signals. andLi,J.Z., editor. *PLoS Comput. Biol.*, **16**, e1008452.
11. Avila Cobos,F., Alquicira-Hernandez,J., Powell,J.E., Mestdagh,P. and De Preter,K. (2020) Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.*, **11**, 5650.
12. Park,S.Y., Gönen,M., Kim,H.J., Michor,F. and Polyak,K. (2010) Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *J. Clin. Invest.*, **120**, 636–644.
13. Fisher,R., Pusztai,L. and Swanton,C. (2013) Cancer heterogeneity: implications for targeted therapeutics. *Br. J. Cancer*, **108**, 479–485.
14. Racle,J., de Jonge,K., Baumgaertner,P., Speiser,D.E. and Gfeller,D. (2017) Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife*, **6**, e26476.
15. Fares,J., Fares,M.Y., Khachfe,H.H., Salhab,H.A. and Fares,Y. (2020) Molecular principles of metastasis: a hallmark of cancer revisited. *Signal Transduct. Target Ther.*, **5**, 28.
16. McDavid,A., Finak,G., Chattopadyay,P.K., Dominguez,M., Lamoreaux,L., Ma,S.S., Roederer,M. and Gottardo,R. (2013) Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinforma. Oxf. Engl.*, **29**, 461–467.
17. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M., Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M., *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.
18. Satija,R., Farrell,J.A., Gennert,D., Schier,A.F. and Regev,A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.
19. Kim,N., Kim,H.K., Lee,K., Hong,Y., Cho,J.H., Choi,J.W., Lee,J.I., Suh,Y.L., Ku,B.M., Eum,H.H., *et al.* (2020) Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.*, **11**, 2285.
20. Bradl,M. and Lassmann,H. (2010) Oligodendrocytes: biology and pathology. *Acta Neuropathol. (Berl)*, **119**, 37–53.
21. Wu,S.Z., Al-Eryani,G., Roden,D.L., Junankar,S., Harvey,K., Andersson,A., Thennavan,A., Wang,C., Torpy,J.R., Bartonicek,N., *et al.* (2021) A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.*, **53**, 1334–1347.
22. Feng,Y., Spezia,M., Huang,S., Yuan,C., Zeng,Z., Zhang,L., Ji,X., Liu,W., Huang,B., Luo,W., *et al.* (2018) Breast cancer development and progression: risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes Dis.*, **5**, 77–106.
23. Cosgrove,N., Varešlija,D., Keelan,S., Elangovan,A., Atkinson,J.M., Cocchiglia,S., Bane,F.T., Singh,V., Furney,S., Hu,C., *et al.* (2022) Mapping molecular subtype specific alterations in breast cancer brain metastases identifies clinically relevant vulnerabilities. *Nat. Commun.*, **13**, 514.
24. Fernández-Nogueira,P., Fuster,G., Gutierrez-Uzquiza,Á., Gascón,P., Carbó,N. and Bragado,P. (2021) Cancer-associated fibroblasts in breast Cancer treatment response and metastasis. *Cancers*, **13**, 3146.
25. Wang,Z., Liu,J., Huang,H., Ye,M., Li,X., Wu,R., Liu,H. and Song,Y. (2021) Metastasis-associated fibroblasts: an emerging target for metastatic cancer. *Biomark. Res.*, **9**, 47.
26. Schulz,M., Salamero-Boix,A., Niesel,K., Alekseeva,T. and Sevenich,L. (2019) Microenvironmental regulation of tumor progression and therapeutic response in brain metastasis. *Front. Immunol.*, **10**, 1713.
27. El Bairi,K., Haynes,H.R., Blackley,E., Fineberg,S., Shear,J., Turner,S., de Freitas,J.R., Sur,D., Amendola,L.C., Gharib,M., *et al.* (2021) The tale of TILs in breast cancer: a report from The International Immuno-Oncology Biomarker Working Group. *Npj Breast Cancer*, **7**, 150.
28. Yin,L., Duan,J.J., Bian,X.W. and cang,Y.S.. (2020) Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Res.*, **22**, 61.
29. Lehmann,B.D., Jovanović,B., Chen,X., Estrada,M.V., Johnson,K.N., Shyr,Y., Moses,H.L., Sanders,M.E. and Pietenpol,J.A. (2016) Refinement of triple-negative breast cancer molecular subtypes: implications for neoadjuvant chemotherapy selection. *PLoS One*, **11**, e0157368.
30. Yao,J., Li,S. and Wang,X.. (2021) Identification of breast cancer immune subtypes by analyzing bulk tumor and single cell transcriptomes. *Front. Cell Dev. Biol.*, **9**, 781848.
31. Bayraktar,S., Batoo,S., Okuno,S. and Glück,S. (2019) Immunotherapy in breast cancer. *J Carcinog.*, **18**, 2.
32. Hanamura,T., Kitano,S., Kagamu,H., Yamashita,M., Terao,M., Okamura,T., Kumaki,N., Hozumi,K., Iwamoto,T., Honda,C., *et al.* (2023) Expression of hormone receptors is associated with specific immunological profiles of the breast cancer microenvironment. *Breast Cancer Res.*, **25**, 13.
33. Vatandoust,S., Price,T.J. and Karapetis,C.S. (2015) Colorectal cancer: metastases to a single organ. *World J. Gastroenterol.*, **21**, 11767–11776.
34. Martin,J., Petrillo,A., Smyth,E.C., Shaida,N., Khwaja,S., Cheow,H.K., Duckworth,A., Heister,P., Praseedom,R., Jah,A., *et al.* (2020) Colorectal liver metastases: current management and future perspectives. *World J. Clin. Oncol.*, **11**, 761–808.
35. Pelka,K., Hofree,M., Chen,J.H., Sarkizova,S., Pirl,J.D., Jorgji,V., Bejnood,A., Dionne,D., Ge,W.H., Xu,K.H., *et al.* (2021) Spatially organized multicellular immune hubs in human colorectal cancer. *Cell*, **184**, 4734–4752.
36. MacParland,S.A., Liu,J.C., Ma,X.Z., Innes,B.T., Bartczak,A.M., Gage,B.K., Manuel,J., Khuu,N., Echeverri,J., Linares,I., *et al.* (2018) Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.*, **9**, 4383.
37. Gong,J., Tu,W., Liu,J. and Tian,D. (2023) Hepatocytes: a key role in liver inflammation. *Front. Immunol.*, **13**, 1083780.
38. Tie,Y., Tang,F., quan,W.Y. and wei,W.X.. (2022) Immunosuppressive cells in cancer: mechanisms and potential therapeutic targets. *J. Hematol. Oncol.*, **15**, 61.
39. Yi,J.S., Cox,M.A. and Zajac,A.J. (2010) T-cell exhaustion: characteristics, causes and conversion. *Immunology*, **129**, 474–481.

40. Yachida,S., Jones,S., Bozic,I., Antal,T., Leary,R., Fu,B., Kamiyama,M., Hruban,R.H., Eshleman,J.R., Nowak,M.A., *et al.* (2010) Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, **467**, 1114–1117.
41. Adamska,A., Domenichini,A. and Falasca,M. (2017) Pancreatic ductal adenocarcinoma: current and evolving therapies. *Int. J. Mol. Sci.*, **18**, 1338.
42. Li,C., Heidt,D.G., Dalerba,P., Burant,C.F., Zhang,L., Adsay,V., Wicha,M., Clarke,M.F. and Simeone,D.M. (2007) Identification of pancreatic cancer stem cells. *Cancer Res.*, **67**, 1030–1037.
43. Wang,L., Liu,Y., Dai,Y., Tang,X., Yin,T., Wang,C., Wang,T., Dong,L., Shi,M., Qin,J., *et al.* (2023) Single-cell RNA-seq analysis reveals BHLHE40-driven pro-tumour neutrophils with hyperactivated glycolysis in pancreatic tumour microenvironment. *Gut*, **72**, 958–971.
44. Peng,J., Sun,B.F., Chen,C.Y., Zhou,J.Y., Chen,Y.S., Chen,H., Liu,L., Huang,D., Jiang,J., Cui,G.S., *et al.* (2019) Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res.*, **29**, 725–738.
45. Yang,J., Lin,P., Yang,M., Liu,W., Fu,X., Liu,D., Tao,L., Huo,Y., Zhang,J., Hua,R., *et al.* (2021) Integrated genomic and transcriptomic analysis reveals unique characteristics of hepatic metastases and pro-metastatic role of complement C1q in pancreatic ductal adenocarcinoma. *Genome Biol.*, **22**, 4.
46. Husain,S. and Thrower,E. (2009) Molecular and cellular regulation of pancreatic acinar cell function. *Curr. Opin. Gastroenterol.*, **25**, 466–471.
47. Zhang,C.Y., Yuan,W.G., He,P., Lei,J.H. and Wang,C.X. (2016) Liver fibrosis and hepatic stellate cells: etiology, pathological hallmarks and therapeutic targets. *World J. Gastroenterol.*, **22**, 10512–10522.
48. Fujita,T. and Narumiya,S. (2016) Roles of hepatic stellate cells in liver inflammation: a new perspective. *Inflamm. Regen.*, **36**, 1.
49. Fernández,L.P., Gómez de Cedrón,M. and Ramírez de Molina,A. (2020) Alterations of lipid metabolism in cancer: implications in prognosis and treatment. *Front. Oncol.*, **10**, 577420.
50. Drew,J. and Machesky,L.M. (2021) The liver metastatic niche: modelling the extracellular matrix in metastasis. *Dis. Model Mech.*, **14**, dmm048801.
51. Sevenich,L. and Joyce,J.A. (2014) Pericellular proteolysis in cancer. *Genes Dev.*, **28**, 2331–2347.
52. Li,S., Zeng,W., Ni,X., Liu,Q., Li,W., Stackpole,M.L., Zhou,Y., Gower,A., Krysan,K., Ahuja,P., *et al.* (2023) Comprehensive tissue deconvolution of cell-free DNA by deep learning for disease diagnosis and monitoring. *Proc. Natl Acad. Sci. U.S.A.*, **120**, e2305236120.
53. Menden,K., Marouf,M., Oller,S., Dalmia,A., Magruder,D.S., Kloiber,K., Heutink,P. and Bonn,S. (2020) Deep learning–based cell composition analysis from tissue expression profiles. *Sci. Adv.*, **6**, eaba2619.
54. Tai,A.S., Tseng,G.C. and Hsieh,W.P. (2021) BayICE: a Bayesian hierarchical model for semireference-based deconvolution of bulk transcriptomic data. *Ann. Appl. Stat.*, **15**, 391–411.

Received: July 27, 2023. **Revised:** November 1, 2023. **Editorial Decision:** December 11, 2023. **Accepted:** December 13, 2023