CDState: an unsupervised approach to predict malignant cell heterogeneity in tumor bulk RNA-sequencing data

Agnieszka Kraft[1,2,3], Josephine Yates[1,3,4], Florian Barkmann[1,3,4], Valentina Boeva[1,3,4]

1 Institute for Machine Learning, Department of Computer Science, ETH Zurich, Switzerland
2 Department of Thoracic Surgery, University Hospital Zurich, Switzerland
3 Swiss Institute for Bioinformatics, Switzerland
4 ETH AI Center, ETH Zurich, Switzerland

Abstract

Intratumor transcriptional heterogeneity (ITTH) presents a major challenge in cancer treatment, particularly due to limited understanding of the diverse malignant cell populations and their relationship to therapy resistance. While single-cell sequencing has provided valuable insights into tumor composition, its high cost and technical complexity limits its use for large-scale tumor screening. In contrast, several databases collecting bulk RNA sequencing (RNA-seq) data from multiple samples across various cancer types are available and could be used to profile ITTH. Several deconvolution approaches have been developed to infer cellular composition from such data. However, most of these methods rely on predefined markers or single cell reference datasets, limiting the performance of such methods by the quality of used reference data. Although unsupervised approaches do not face such limitations, existing methods have not been specifically adapted to characterize malignant cell states, and focus instead on general cell types. To address these gaps, we introduce CDState, an unsupervised method for inferring malignant cell subpopulations from bulk RNA-seq data. CDState utilizes a Nonnegative Matrix Factorization model improved with sum-to-one constraints and a cosine similarity-based optimization to deconvolve bulk gene expression into distinct cell state-specific profiles, and estimate the abundance of each state across tumor samples. We validate CDState using bulkified single-cell RNA-seq data from five cancer types, showing that it outperforms existing unsupervised deconvolution methods in both cell state proportions and gene expression estimation. Applying CDState to 33 cancer types from TCGA, we identified recurrent malignant cell programs, with epithelial-mesenchymal transition as the main driver of tumor transcriptional heterogeneity. We further link the identified malignant states to patient clinical features, revealing states associated with worse patient prognosis. We show that ITTH is linked with patient survival

and clinical features, as well as is associated with varied responses to therapy. Finally, we identified potential genetic drivers of ITTH and malignant cell states, including *TP53*, *KRAS*, and *PIK3CA*, known oncogenic genes.

Introduction

Single-cell sequencing technology has significantly advanced our understanding of intratumor transcriptional heterogeneity (ITTH), revealing the diverse gene expression programs within tumor-resident cells. This heterogeneity has been shown to be implicated in cancer progression, metastasis, and therapy resistance [1–3]. While pioneering studies focused specifically on the tumor microenvironment (TME) composition, recently more emphasis has been put on the heterogeneity within the malignant compartment [4–9]. For instance, single-cell transcriptomic analyses of glioblastoma, melanoma, head and neck squamous cell carcinoma, and colorectal cancer have identified distinct malignant cell states, characterized by differences in differentiation status, stress response, and plasticity [4,5,7,10]. These states are closely intertwined with TME composition, influencing processes such as angiogenesis [11], and immune evasion [7,12,13]. Notably, pan-cancer studies have suggested that recurrent malignant cell states are shared across multiple tumor types, implying that common oncogenic programs may drive ITTH across cancers [14,15]. These findings raise the possibility that targeting specific malignant states could offer a broadly applicable therapeutic strategy.

Unfortunately, single-cell sequencing remains costly and technically challenging, making it impractical for comprehensive screening of big tumor datasets. Conversely, multiple databases with bulk RNA sequencing (RNA-seq) data profiled across many samples and cancer types, provide a promising alternative for robust characterization of tumor composition. Over the past decade, numerous computational deconvolution methods have been developed to infer cell type proportions from bulk RNA-seq data [16–20]. However, most of these methods rely on reference transcriptomic data for cell types present within the bulk mixture, requiring a priori knowledge of the sample composition and often assuming a stable distribution of cell types across samples. More critically, existing methods are unable to discover malignant cell states absent in the reference data, limiting their ability to accurately resolve tumor-specific transcriptional heterogeneity.

To address these limitations, we introduce CDState, an unsupervised deconvolution framework based on Nonnegative Matrix Factorization (NMF). Unlike conventional deconvolution approaches, CDState is designed to identify and characterize malignant cell states in bulk RNA-seq data without requiring predefined marker genes or reference

datasets. To ensure biologically meaningful deconvolution, CDState integrates sum-to-one constraints on the estimated proportions, and a cosine similarity-based optimization term, which encourages the disentanglement of cell states with similar gene expression profiles.

We validate CDState using bulkified single-cell RNA-seq datasets from glioblastoma [5], lung adenocarcinoma [21], breast cancer [22], ovarian high grade serous carcinoma [23], and squamous cell carcinoma [24], showing that it outperforms existing unsupervised deconvolution methods [25–30] in both cell state proportions and gene expression estimates across diverse tumor contexts.

To provide a comprehensive characterization of ITTH across cancers, we apply CDState to 33 cancer types from The Cancer Genome Atlas (TCGA). We show that ITTH is associated with patient clinical features, including survival, molecular subtype and tumor stage. Moreover, we found it to be linked with a varied response to therapy. We uncover recurrent gene expression programs driving ITTH across cancers, with epithelial-mesenchymal transition emerging as dominant transcriptional driver. In addition, we show that mutations in genes such as *TP53* or *KRAS*, are significantly associated with ITTH and proportions of malignant states, suggesting potential genetic mechanisms underlying transcriptional heterogeneity.

Our work provides novel insights on transcriptional intratumor heterogeneity across multiple tumors, highlighting potential new molecular targets and biomarkers for future treatment and patient stratification.

Results

Overview of CDState

CDState is a framework based on Nonnegative Matrix Factorization (NMF), with additional constraints to improve the accuracy and interpretability of deconvolution (Figure 1A-B). The inputs to CDState are normalized bulk RNA-seq data and tumor purity estimates.

Compared to other NMF-based methods which rescale weights after deconvolution, CDState imposes sum-to-one constraints on estimated weights into the optimization process (Figure 1B, Methods). In addition, the objective function includes an additional term that penalizes high cosine similarity between deconvolved source gene expression profiles. This encourages greater distinction between transcriptionally similar states, for example between fibroblasts and malignant cells with mesenchymal properties.

The relative influence of the two optimization terms - reconstruction error and source transcriptional profile separation - is controlled by two parameters, α and β, which sum to 1 (Figure 1B, Methods). Overall, CDState consists of two steps. In the first step, CDState minimizes reconstruction error (β = 0). In the second step, β is gradually increased, giving more importance to minimizing cosine similarity of source gene expression profiles, while temporarily allowing for higher reconstruction error. The selection of values for α and β is guided by monitoring the mean squared error (MSE) between the estimated and true purity of deconvolved samples (Methods). Once the MSE stops decreasing, β is gradually reduced back to zero for a final optimization step that minimizes reconstruction error (Figure 1B, Methods).

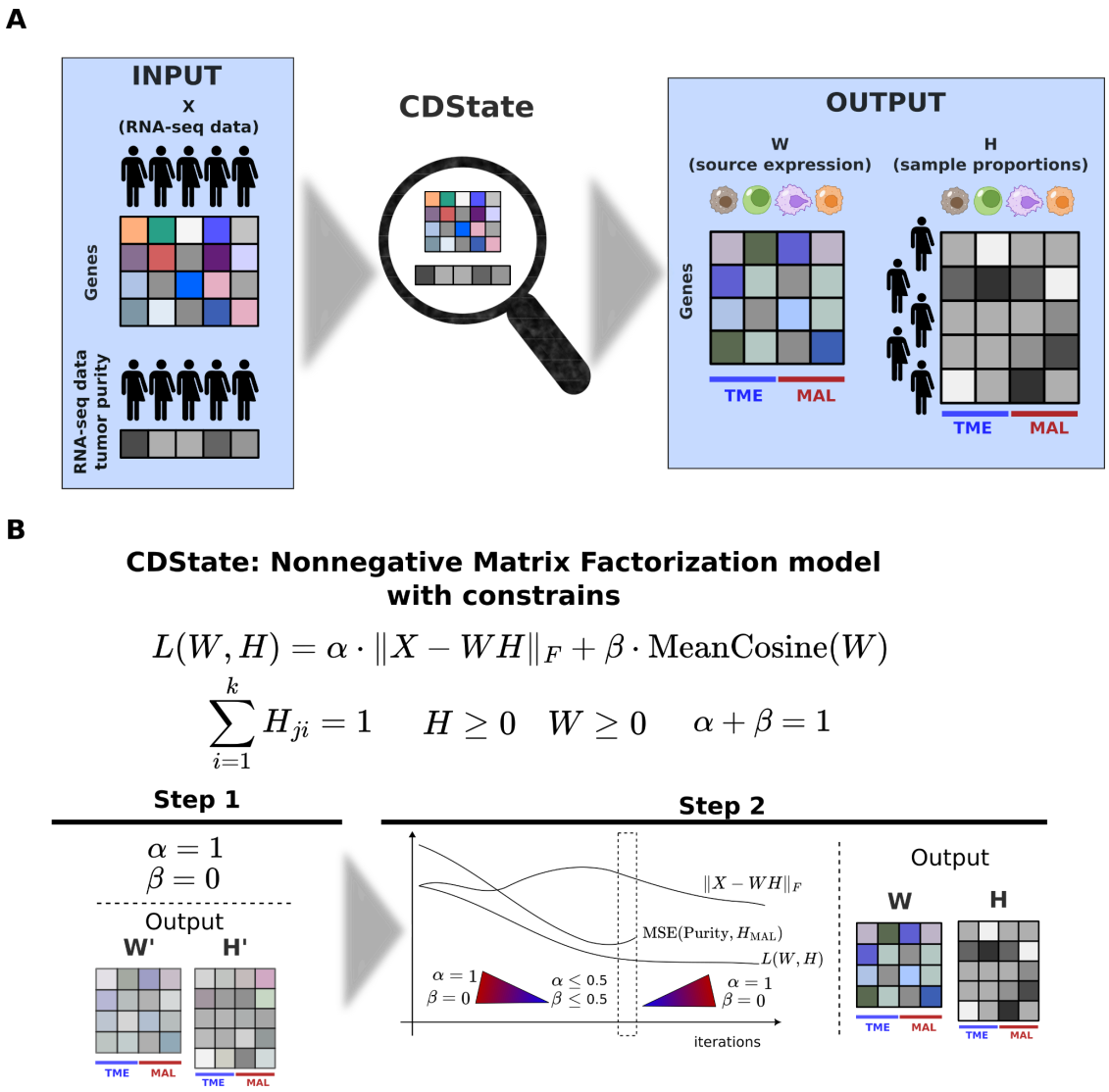

Figure 1. CDState overview. A. CDState takes as input a normalized RNA-seq expression matrix along with sample purity estimates. As output it returns inferred proportion and source expression

matrices, along with source assignment to tumor microenvironment (TME) or malignant (MAL) cell states. **B.** CDState algorithmic flow and parameters.

CDState outperforms existing methods in deconvolving cell type proportions and gene expression

We first tested CDState on bulkified single-cell RNA sequencing data from lung adenocarcinoma (*Kim et al.* [21]). Specifically, we created four synthetic bulk RNA-seq mixtures using three cell types - malignant cells, B cells, and fibroblasts, for which cell-type proportions were sampled from Dirichlet distributions with varying alpha parameters (Figure 2A, Methods). With this procedure, we assess CDState's performance under different conditions, including mixtures where pure samples of specific cell types were available (MixA and MixC), certain cell types were present in very low proportions (MixD), or no pure samples were included (MixB).

We compared CDState against four state-of-the-art unsupervised deconvolution methods: Archetype Analysis (AA) [29,30], debCAM [27,28], Linseed [26], and CellDistinguisher [25]. These methods were selected based on a recent benchmark highlighting them as top-performing deconvolution tools [31], while AA was included as it was previously used to study molecular phenotypes in cancer [32]. To account for variability in inferred solutions, all methods (except for the deterministic CellDistinguisher) were run 20 times with different random seeds. To ensure a fair comparison, we set all methods to deconvolve the data into three components, matching the known number of cell types in the mixtures. Additionally, we included results from CDState's first optimization step (Figure 2B, CW-NMF) to verify the impact of the cosine similarity term.

CDState consistently achieved the highest correlation and lowest root mean squared error (RMSE) with both ground truth source profiles and proportions (Figure 2A, Suppl. Figure S1). Notably, it showed the smallest variability in performance metrics across multiple runs, indicating robust and stable results. Furthermore, the cosine similarity-based optimization improved CDState performance, by reducing RMSE and improving correlation with true source profiles compared to the CW-NMF results (Figure 2B, Suppl. Figure S1). This improvement was especially evident in more complex settings, such as MixB and MixD, where no pure samples were available. As expected, we also observed a decrease in cosine similarity between the inferred and ground truth sources, further validating the role of our cosine similarity term (Suppl. Figure S2).

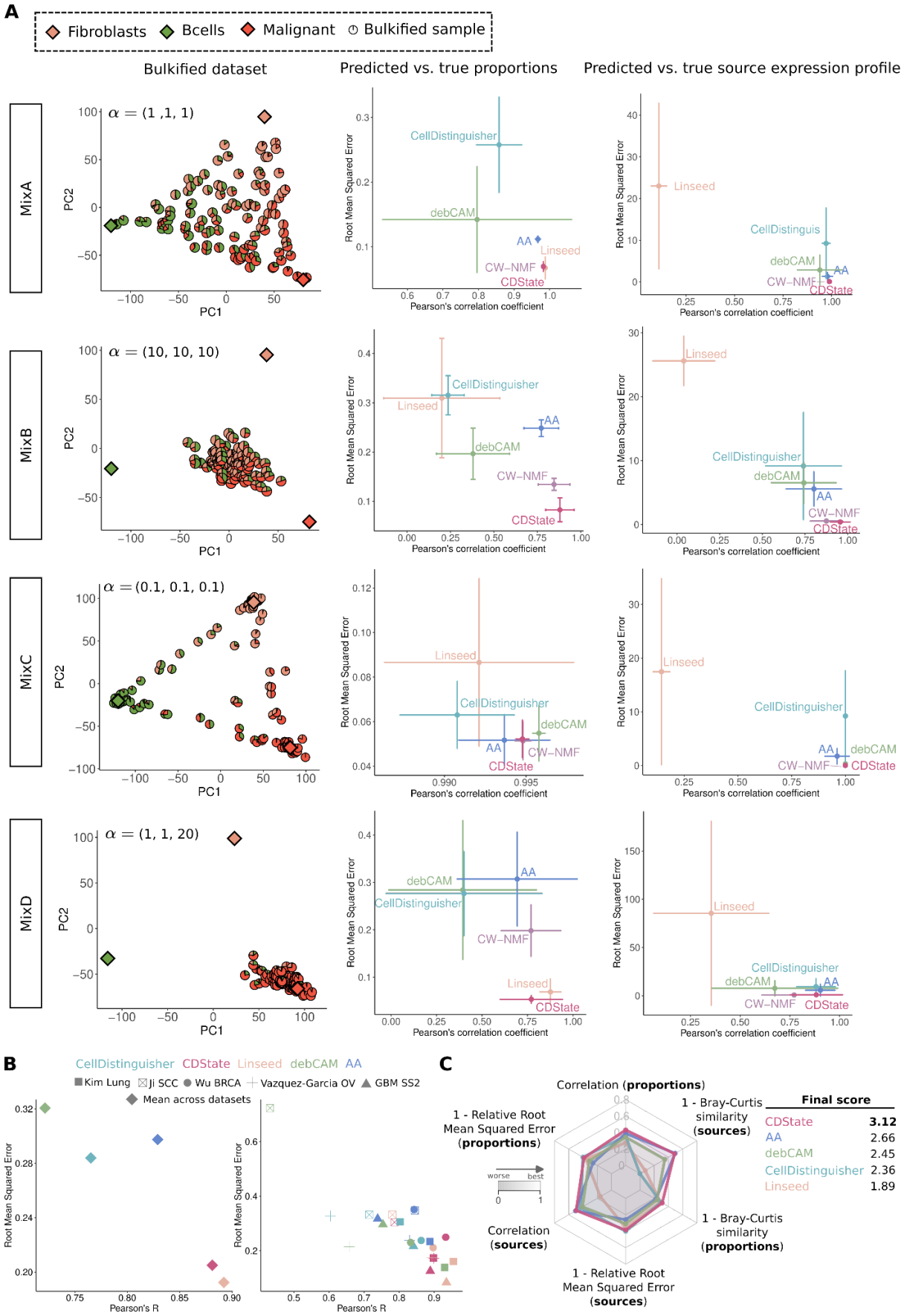

Figure 2. CDState validation. A. Deconvolution performance on four synthetic mixtures (in rows), using four external methods and the first step of CDState (CW-NMF). Mixtures were generated by

sampling cell type proportions from four Dirichlet distributions with varying alpha parameters. These proportions were used to combine malignant cells, B cells, and fibroblast from the *Kim et al.* dataset (Methods). Bulkified samples were visualized in the PCA space projected on the ground truth source gene expression. Pie chart coloring corresponds to the proportion of respective cell type in each bulkified sample (left column). Deconvolution performance was compared using Root Mean Squared Error and Pearson's correlation coefficient on both inferred proportions (middle column) and source gene expression profiles (right column). **B**. Comparison of deconvolved and true tumor purity across five bulkified cancer datasets (left) and per dataset (right). Point shapes correspond to the dataset, while colors correspond to deconvolution methods. **C.** Performance summary across the five bulkified cancer datasets. Each metric ranges from 0 to 1, with values close to 1 being the best. The final score was calculated as the sum of individual metrics.

CDState outperforms existing methods in deconvolving malignant cell state proportions and gene expression

To evaluate CDState's performance in deconvolving malignant cell states, we applied it to five bulkified single-cell RNA-seq datasets from breast cancer [22], ovarian cancer [23], glioblastoma [5], lung cancer [21], and squamous cell carcinoma [24] (Methods).

We first assessed each method's ability to distinguish malignant cells from tumor microenvironment (TME). CDState was the second best method, achieving high correlation with tumor purity and low RMSE in all datasets and the best correlation with tumor purity in breast and ovarian cancer datasets (Figure 2B). Next, we assessed CDState's performance in identifying malignant cell states and their proportions. For each malignant state, we computed Pearson's correlation coefficient, relative RMSE (rRMSE), and Bray-Curtis index (BC) between the true and inferred source gene expression profiles and proportions (Figure 2C, Suppl. Figure S4, Methods). For the overall score, we transformed the rRMSE and Bray-Curtis index, to *1 - rRMSE* and *1 - BC*, respectively, to obtain metrics whose higher scores reflected better performance. Finally, overall performance of each method was calculated as the sum of respective metrics across datasets.

CDState showed overall the best performance compared to existing deconvolution methods (Figure 2C, Suppl. Figure S3, Suppl. Table S4). It achieved the highest overall correlation and BC index with true proportions (R = 0.54, BC = 0.4). For all other metrics, CDState was the second best, consistently showing high performance (Figure 2C, Suppl. Table S4).

Additionally, CDState was among the best performing methods for the number of correctly identified malignant states. Specifically, it correctly separated 28 out of 30 states based on

correlation between true and inferred proportions, and all of the ground truth states based on the correlation with source gene expression (Suppl. Figure S3).

CDState accurately infers gene expression and proportions of glioblastoma malignant cell states

To assess CDState's performance in a realistic setting in which the true number of states is unknown, we applied it to the bulkified glioblastoma (GBM) dataset [5].

We determined the optimal number of states by selecting the number of sources (k) that maximized the mean correlation with tumor purity across 20 runs. This approach led to selecting a seven-state solution (Suppl. Figure S4). After identifying malignant and TME components, we evaluated their correlation with the ground truth proportions of corresponding cell types (Figure 3A). CDState-inferred proportions were highly correlated with ground truth cell type proportions, correctly separating malignant and non-malignant compartments. When comparing correlations for specific malignant states, we found a clear separation of MESlike1 and NPClike states. In contrast, MESlike2 and AClike states proportions correlated with the same inferred state S1, while OPClike, NPClike2 and cycling cell proportions all correlated with state S5 (Figure 3A). The difficulty in resolving these states can be explained by the strong correlation of their true proportions in bulk mixtures, particularly evident for OPClike, NPClike2 and cycling cells (Suppl. Figure S5). Although the selected solution did not distinguish between these particular states, analysis of CDState solutions with more states showed their more refined separation (Suppl. Figure S5-S6).

To further validate observed correlations between true and inferred proportions, we selected marker genes of CDState-identified states and scored single cells composing the bulkified dataset. The obtained scores reflected identified associations with proportions, further confirming CDState performance (Figure 3B). Additionally, we mapped these single-cell scores onto the transcriptional space of the original *Neftel et al.* [5] malignant cells, to better visualize the distribution of scores across cells in different states. CDState successfully identified the main axes of variation, with two OPC-NPC gradients (S5 one stronger for OPC-like cells, S4 for NPC-like cells) and two AC-MES gradients (S6 stronger for AC-like cells, and S1 for MES-like cells) (Figure 3C).

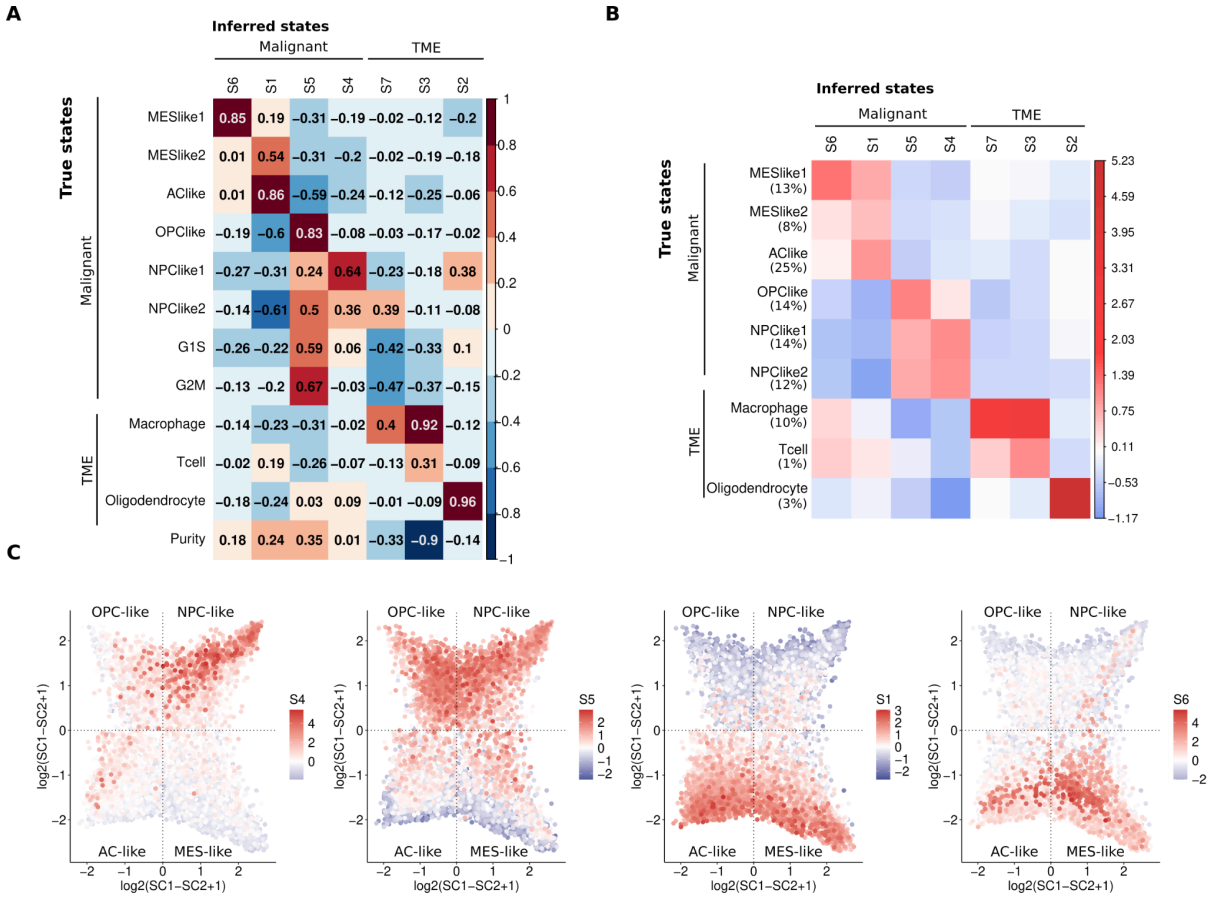

Figure 3. CDState deconvolution of cell states in bulkified glioblastoma data. A. Heatmap with Pearson's correlation coefficient between true and CDState estimated proportions. **B.** Single cell scores of CDState deconvolved states. Scores were calculated using ANS method (*Ciernik et al.*) [33], and z-score normalized. Percentages represent the proportion of cells in the bulkified dataset. **C.** Two-dimensional representation of glioblastoma malignant cell states from *Neftel et al.* [5] Each quadrant corresponds to one state (OPC-like, NPC-like, AC-like, MES-like). Position of the malignant cells was determined by the relative scores of the cell states, and colors reflect the scores of CDState identified malignant states.

Pan-cancer inference of malignant cell states using TCGA bulk RNA-seq data

The primary goal of CDState is its application to bulk RNA-seq data from large patient cohorts to identify robust states that drive malignant cell heterogeneity. To achieve this goal, we applied CDState to 33 datasets combining 9,669 patient samples from The Cancer Genome Atlas (TCGA), to provide a pan-cancer characterization of malignant cell states (Figure 4).

Overall, CDState achieved good separation of malignant and non-malignant compartments, with 27 out of 33 TCGA cancer datasets showing a correlation of at least 0.5 between estimated and true tumor purity (Suppl. Figure S8). Additionally, in 31 of 33 cancer types, deconvolved malignant gene expression showed stronger association with copy number alterations compared to bulk RNA-seq data. This observation indicates that the deconvolved expression more accurately reflects the true malignant compartment, whose expression is typically linked to the sample copy number profiles (Suppl. Figure S8).

Next, we used the inferred proportions of identified states to assess sample-specific transcriptional heterogeneity. Specifically, we calculated Intratumor Transcriptional Heterogeneity score (ITTH) using these proportions and the formula for Shannon's entropy index (Methods). TCGA samples showed a varied distribution of ITTH score, with the lowest median score in thymoma (THYM), and the highest in adrenocortical carcinoma (ACC) (Figure 4A).

To characterize the molecular programs driving specific malignant states, we tested the enrichment of state-specific marker genes in 50 hallmarks of cancer gene sets deposited in MSigDB (Methods). Overall, we observed recurrent enrichment of molecular programs within and across cancer types, with epithelial-mesenchymal transition as the main driver of malignant cell states, enriched in 132 out of 193 identified states (Figure 4C, Suppl. Table S8). Surprisingly, we found enrichment of apoptosis and P53 pathways as the second and third most frequently enriched pathways across the states. Among other most recurrent programs, we found pathways linked to immune response, differentiation, metabolism, and hormone-regulation (Figure 4C, Suppl. Table S7-S8). These findings suggest that while each cancer type expresses distinct malignant states, certain core transcriptional programs are conserved across malignancies.

Malignant cell states are associated with patient clinical features

To assess the clinical significance of the inferred malignant cell states, we verified their association with patient clinical features. The ITTH index was significantly associated with overall survival in five cancer types (Figure 4B, Suppl. Table S6, p-value < 0.05). Interestingly, in clear cell renal carcinoma (KIRC), head and neck squamous cell carcinoma (HNSC), and breast cancer (BRCA), higher scores correlated with worse patient prognosis. Conversely, in liver hepatocellular carcinoma (LIHC) and uterine carcinoma (UCEC), higher heterogeneity was linked to a lower hazard.

At the malignant cell state level, malignant states in 17 cancer types were significantly associated with worse prognosis, while states in 10 cancer types were significantly linked to better survival (Figure 4D, Suppl. Table S6). Functionally, states associated with poor prognosis were predominantly enriched in pathways related to development and differentiation, as well as immune response and metabolism (Suppl. Table S7-S8).

Further, ITTH score was significantly associated with tumor molecular subtypes in 14 cancer types, and patient clinical features in 9 cancer types (Figure 4E-F). Interestingly, we found higher ITTH linked with worse patient response to chemotherapy and targeted molecular treatment in four cancer types (Figure 4G).

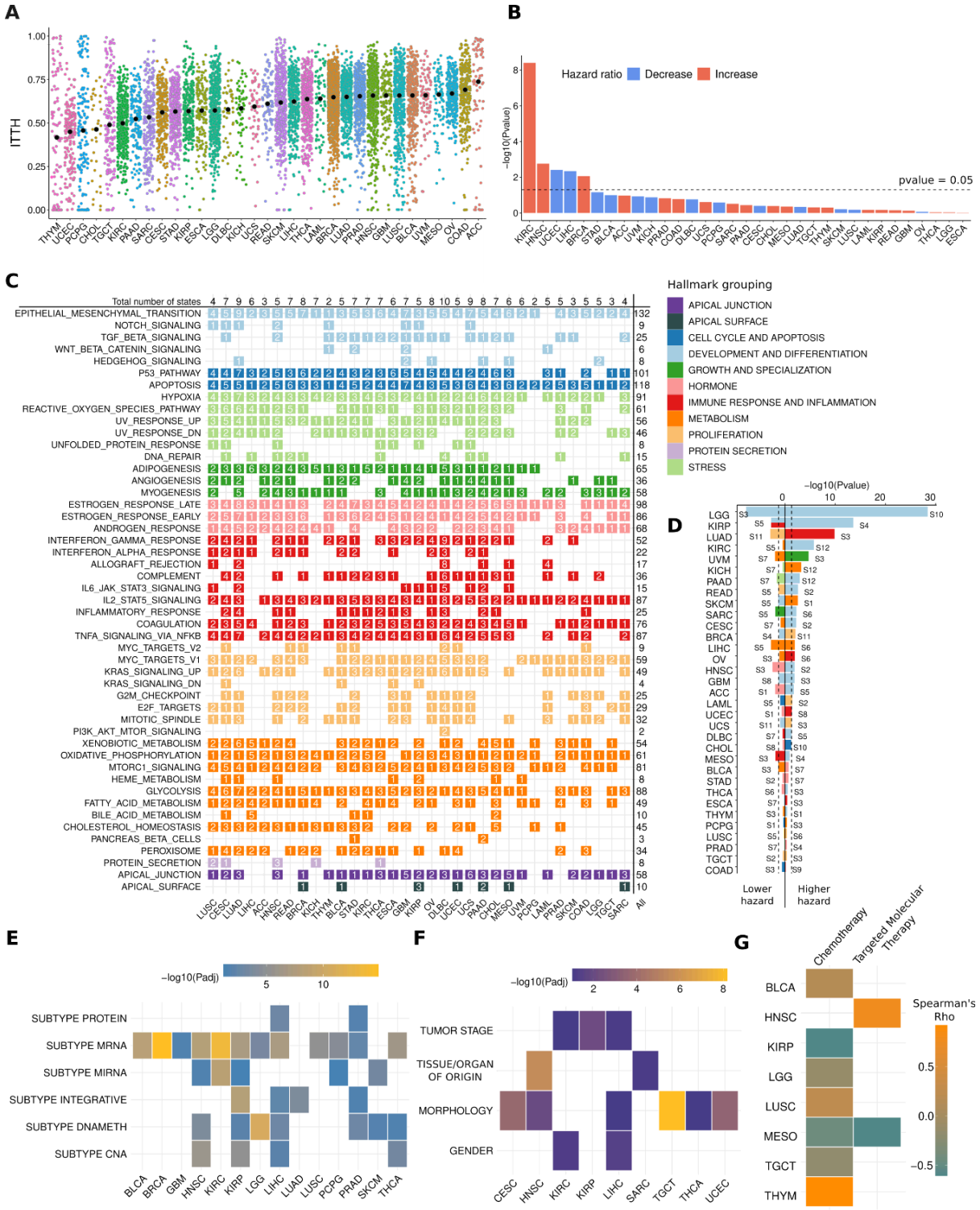

Figure 4. Pancancer characterization of malignant cell heterogeneity using CDState. A. TCGA cancer types present a varied degree of intratumor transcriptional heterogeneity (ITTH), calculated as Shannon entropy of deconvolved proportions. **B.** Association between ITTH and patient survival across TCGA datasets. P-value was calculated using Cox univariate regression analysis. **C.** Hallmarks of cancer enriched in the deconvolved malignant cell states across TCGA datasets. Numbers correspond to the number of states per cancer type in which the hallmark was significantly (p-value < 0.05) enriched. Colors correspond to the hallmark grouping based on the shared molecular

process. The enrichment was calculated using a hypergeometric test. **D.** States with significant association with TCGA patient survival. Colors correspond to the group of the most enriched hallmark in the given state. The survival association was evaluated using Cox univariate regression analysis. Dotted lines represent p-value = 0.05. **E-F.** Heatmaps show significant (Kruskal-Wallis adjusted p-value < 0.05) associations between ITTH and tumor molecular subtypes (**E**) and patient clinical characteristics (**F**). **G.** Association between ITTH and response to chemotherapy across TCGA cancer types (Kruskal-Wallis p-value < 0.05).

Genetic drivers of intratumor transcriptional plasticity

To identify potential genetic drivers of ITTH and malignant states, we used point biserial correlation to find genes with mutations or copy number changes most associated with our heterogeneity scores or state proportions. In case of copy number alterations, we focused on identifying gene losses and gains in transcriptional factors, given their prominent role in gene expression regulation.

We found significant association between ITTH and gene mutations in 13 cancer types, and copy number changes in transcriptional factors in 20 cancer types (Figure 5A, Suppl. Figure S9A). Among mutations, most of them were associated with higher ITTH score, but in 4 cancer types, they were linked to lower heterogeneity. Several known cancer-related genes showed significant association, including *KRAS*, *PTEN*, *CTNNB1*, *TP53*, and *PIK3CA* (Figure 5A). In case of association with copy number changes, we observed gene copy losses more often than gains (Suppl. Figure S9A). Among associated genes, we found several genes previously linked to cancer, including *NFE2L2, MEIS2*, and *KLF13*.

Next, we analyzed events most associated with proportions of individual malignant cell states and assessed prevalence of affected genes across analyzed cancer types. We found mutations in *TP53* to be the most commonly associated with malignant cell state proportions, followed by mutations in *KRAS*, *TTN* and *BRAF* (Figure 5B).

Given that our results have shown that gene expression programs are recurrently enriched across malignant cell states, we wanted to verify a similar hypothesis using genetic events. In case of mutated genes, we found mutations in *TP53* to be the most frequently associated mutations with malignant cell states across 12 cancer types, followed by mutations in *BRAF*, *CTNNB1*, *KRAS* and *PIK3CA*, each associated with malignant state proportions in three cancer types (Figure 5C). Among copy-number affected transcriptional factors, the top recurrent genes included *HES2*, *ACOT7* and *MTMR3* (Suppl. Figure 9B).

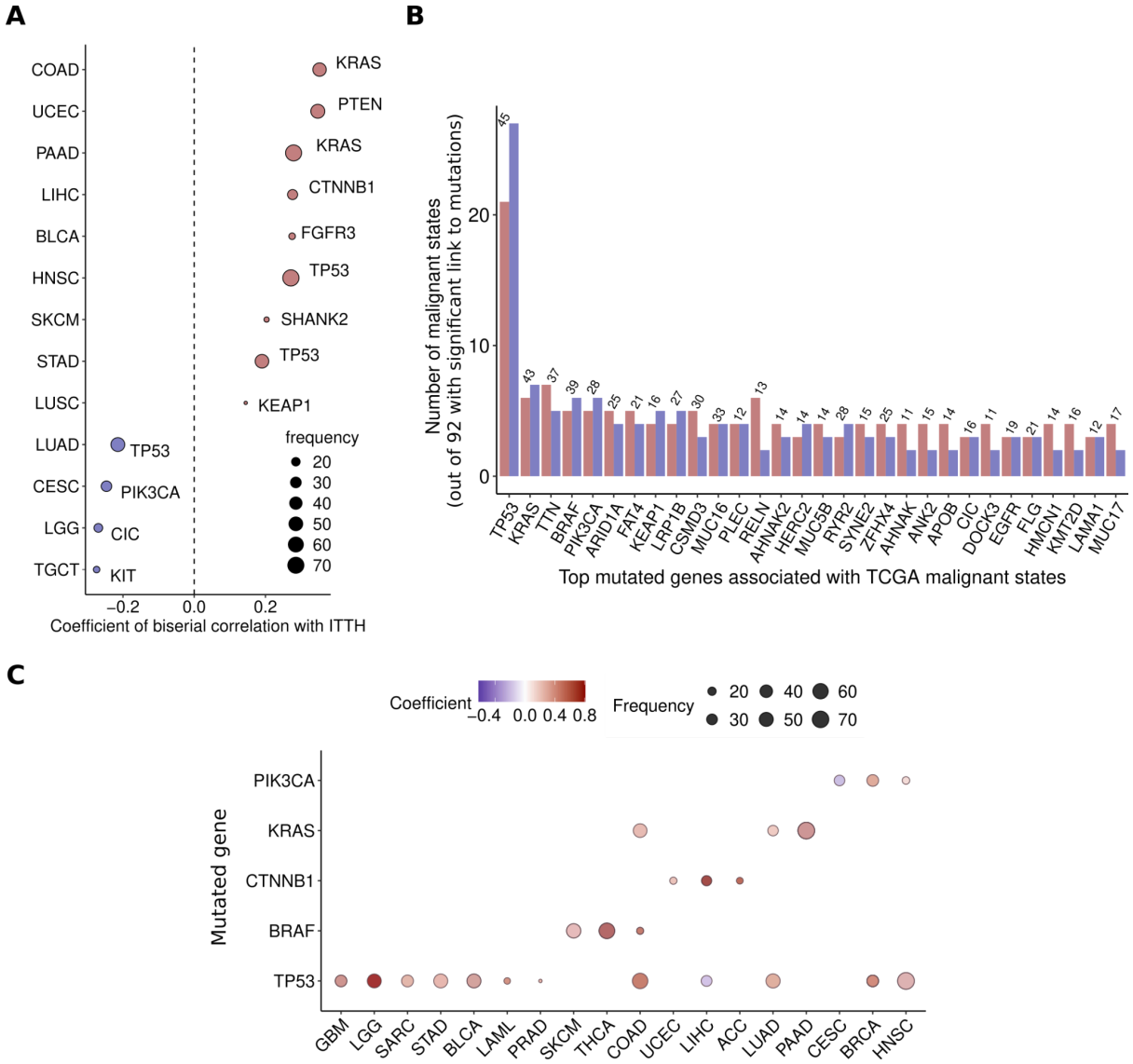

Figure 5. Candidate genetic drivers of TCGA malignant cell heterogeneity. A. Gene mutations with the highest absolute biserial correlation (adjusted p-value < 0.05) with cancer-type specific heterogeneity score (ITTH). Cancer-specific frequency of event is represented by size of points. **B.** Top 30 mutated genes most frequently associated with TCGA malignant cell states (adjusted p-value < 0.05). Numbers on top of bar plots represent the mean frequency [%] of mutations across all analyzed TCGA samples. Blue and red bars represent negative and positive associations, respectively. **C.** Gene mutations with the highest significant association (adjusted p-value < 0.05) with malignant state proportions, recurrently linked to states across TCGA cancer types.

Discussion

Multiple published studies provide evidence on the significance of transcriptional heterogeneity in tumors, particularly in relation to treatment resistance and patient prognosis. With the rapid expansion of single-cell technology and its implementation in studying tumor samples, we have gained a broader understanding of malignant cell transcriptional profiles

across multiple cancer types. However, current single-cell studies do not offer sufficiently large sample sizes to fully characterize transcriptional heterogeneity. At the same time, large cancer atlases, such as TCGA, could be utilized for this task. This, however, requires availability of appropriate methods to separate transcriptional signals coming from different cell states present in bulk data. In this work, we introduce CDState, an unsupervised deconvolution method specifically developed to address this challenge. We demonstrate that CDState outperforms other existing unsupervised methods across bulkified cancer datasets with diverse composition, accurately identifying previously reported malignant cell states and their proportions.

Our analysis of TCGA datasets using CDState identified recurrent malignant cell programs, with epithelial-mesenchymal transition as the key driver of malignant cell heterogeneity across cancer types. We proposed an intratumor transcriptional heterogeneity index (ITTH) calculated based on proportions of inferred malignant cell states, and showed its association with patient survival in several cancer types. Moreover, our ITTH index was significantly linked with patient clinical features, including molecular subtype of disease, tumor stage and response to therapy. These results highlight its potential as a biomarker for patient stratification and survival prediction.

Our results on the association between genetic alterations and specific malignant cell states, show evidence on the interplay between genetic and transcriptional heterogeneity, highlighting the potential role of specific genes in driving malignant cell heterogeneity. These findings may be particularly important for the development of novel diagnostic and therapeutic approaches, especially for cancer types with low survival rates. For instance, we found *TP53* mutation to be the most frequently associated with either higher or lower malignant state proportions. While for now, *TP53* remains untargetable, several clinical trials explore it as a treatment target [34,35]. Our analysis identified several potential genetic drivers, however, further extensive experimental validation using CRISPR-Cas9 and similar approaches should be used to validate these findings..

While we showed the utility and high performance of our approach across varied bulk samples, there are some limitations that should be taken into consideration. Firstly, CDState requires tumor purity estimates which are used to guide identification of malignant cell states and optimization of the parameters. Therefore, the deconvolution output heavily depends on the reliability of the provided tumor purity estimates. Recent study has shown significant discrepancy between purity estimates using DNA and RNA sequencing data making the decision on the use of appropriate approach challenging [36,37]. In our analysis of the TCGA data, we used tumor purity estimates obtained using matched DNA-sequencing data and

ABSOLUTE [38]. Therefore, given the purity estimates were obtained from different pieces of tumors than the analyzed bulk RNA-seq data, we acknowledge this could potentially affect our deconvolution results. Secondly, it is challenging to systematically decide on the optimal number of components for deconvolution. We showed that increasing the number of deconvolved cell types leads to improved separation of sources. However, selection of too many states may lead to identification of sample-specific sources.

Despite these challenges, CDState is a promising alternative to supervised deconvolution approaches for big cancer atlases with bulk data, when accurate single-cell references are unavailable. CDState can be used to identify clinically relevant biomarker signatures that could potentially be used in clinical routine for patient stratification and diagnostic testing. Additionally, it can help identify relevant molecular pathways for future cancer research.

Methods

CDState algorithm

CDState is a computational method designed to identify and enumerate transcriptional cell states within tumor-resident cells. The algorithm operates in two main steps (Figure 1A-B), combining constrained Non-negative Matrix Factorization with an iterative optimization strategy inspired by simulated annealing [39], to maximize the dissimilarity between gene expression profiles of inferred transcriptional states.

Step 1: Identification of initial solution

1) CDState applies NMF to decompose the bulk RNA-sequencing expression matrix into a predefined number of components (i.e., cell states).
2) At this stage, the optimization focuses solely on minimizing the reconstruction error, with alpha (α) set to 1 and beta (β) set to 0.
3) The initial solution includes both malignant and tumor microenvironment (TME) components, which are refined in Step 2.

Step 2: Optimization of cell state separation

1) To improve the separation of inferred transcriptional states, β is progressively increased, weighting the optimization towards minimizing the mean cosine similarity between inferred cell state gene expression vectors.

2) As β increases, the method more likely accepts a higher reconstruction error while expanding the optimization space to maximize cell state dissimilarity.
3) This annealing-inspired process continues until the mean squared error between inferred and true tumor purity stops decreasing, at which point β is gradually reduced back to zero.
4) A final round of optimization, focused on minimizing reconstruction error runs until convergence.

Constrained Nonnegative Matrix Factorization (NMF)

The CDState algorithm assumes that the gene expression of a mixed bulk sample is a linear combination of expression of pure source cell types:

$x_i = \sum_{k=1}^{K} h_k w_{ik}$ given *K* sources, where x_i is bulk gene expression of gene *i*, w_{ik} is gene expression of gene *i* in pure source *k*, and h_k is a proportion of source *k*. Therefore, the identification of underlying sources in the *n* input samples with bulk gene expression values for *m* genes can be formulated as a linear problem. CDState is based on the following Nonnegative Matrix Factorization framework:

$X \approx WH$, where *X*, *W* and *H* are *m* x *n*, *m* x *K* and *K* x *n* non-negative matrices representing the input bulk expression, latent cell states and proportions, respectively.

In this formulation, NMF tends to return a sparse matrix of sources, in which each source represents an independent underlying element of the input data. In a biological setting, these sources can be interpreted as individual biological processes, such as proliferation or stress-response. In reality, a living cell which is in a particular state can express more than one of these processes simultaneously. To account for this overlap, we extend the NMF model with the *sum-to-one* constraint on weights, such that:

$\sum_{k=1}^{K} h_k = 1$ for each sample.

Our CDState algorithm uses the following objective function based on NMF:

$L(W, H) = \alpha||X - WH|| + \beta MeanCosine(W)$, where

$\sum_{k=1}^{K} h_k = 1$, $H, W \geq 0$, $\alpha + \beta = 1$, and $MeanCosine(W)$ is the mean of the pairwise cosine similarities of source gene expression vectors.

Simulated annealing-based optimization

Simulated annealing is incorporated into CDState to balance the trade-off between minimizing reconstruction error and maximizing the biological difference of inferred transcriptional states. Initially, the optimization focuses purely on reconstructing the bulk RNA-seq data by minimizing the reconstruction error (α = 1, β = 0). However, to avoid inference of highly correlated cell states, in the second step of the algorithm CDState gradually increases the β parameter, which places greater emphasis on minimizing the mean cosine similarity between inferred states. By progressively accepting a higher reconstruction error, CDState encourages the discovery of transcriptional states that are more biologically distinct. However, to ensure that the final decomposition remains relevant, β is only increased as long as the mean squared error (MSE) between the inferred and true tumor purity continues to improve. Once MSE begins to rise, β is reduced back to zero, and a final round of reconstruction optimization is performed until convergence.

Implementation details

Inputs:

a) Normalized bulk RNA-seq expression profiles (matrix of shape m genes × n samples).
b) Tumor purity estimates.
c) A predefined range of possible cell state numbers (e.g., “k=2:4” runs CDState for 2, 3, and 4 states).

Outputs:

a) Matrix with gene expression profiles of the identified cell states (m genes × k sources).
b) Matrix with proportions of the identified cell states in the input samples (n samples x k sources).
c) Component assignment to malignant or tumor microenvironment states.

Optimization Strategy:

a) The weights and source update in CDState step 1 is solved using the SCIPY minimize-maximize function with Sequential Least Squares Programming (SLSQP) optimizer and non negative least squares objective function.
b) In step 2, the source matrix is updated using gradient descent (using the JAX python library).

Initialization and multiple runs:

a) Due to the stochastic nature of NMF, CDState is run multiple times with different starting sources.
b) The starting sources are selected from bulk samples using K-means clustering, with the number of clusters corresponding to the selected number of states. A random sample from each cluster is then selected as the starting source for each run.

Gene filtering

Input bulk data is first filtered to minimize the impact of outlier genes and reduce computation time, using the method described in [28]. Specifically, genes with vector norm below 30% or above 99% of the mean vector norm across all genes, are excluded. Within the remaining genes, their expression levels are binned based on 0.1 increase step, and only the top 50% of genes with the highest standard deviation within each bin are retained (Additional File 1). This filtered gene set is then used for the entire inference process. However, a function to infer gene expression for the complete gene set is also available within the CDState package.

Definition of TME and malignant cell states

After the initial step of the algorithm, CDState classifies the inferred sources into tumor microenvironment (TME) and malignant cell states. This is achieved through a greedy search to identify the optimal combination of source proportions that have the strongest Spearman's correlation with the ground truth purity. Specifically, the proportions of each inferred component are correlated with the provided purity estimates and ranked in descending order of correlation strength. Then, for each consecutive subset of components, the aggregated proportions are correlated with true purity. The subset showing the highest correlation (with a one-sided p-value < 0.05) is selected as representing malignant cell states, while the remaining sources are selected as TME components. If purity estimates are available for fewer than 20 samples, the significance of the correlation is assessed using bootstrap resampling with 1,000 iterations.

Selection of the number of cell states for deconvolution

To determine the optimal number of cell states for deconvolution, CDState provides several metrics, including reconstruction error, total error, the assignment of components to malignant and TME states, and the final Pearson's correlation with true purity based on inferred malignant cell state proportions. Users may select the most suitable number of states for deconvolution using these metrics.

Generation of bulkified datasets

To generate bulkified datasets, we followed two distinct approaches: one for cell type deconvolution, and another specifically for malignant cell state inference. For the general cell type mixtures, we selected three distinct cell types from the *Kim et al. [21]* lung adenocarcinoma dataset [21]: malignant cells, B cells and fibroblasts. Then, we used Dirichlet distribution to simulate varying cellular compositions, with the alpha parameter controlling the level of variation. Specifically, we generated four mixtures in which proportions of each bulk sample were drawn from the following distributions:

- MixA $\theta_i \sim Dir(1, 1, 1)$
- MixB $\theta_i \sim Dir(10, 10, 10)$
- MixC $\theta_i \sim Dir(0.1, 0.1, 0.1)$
- MixD $\theta_i \sim Dir(1, 20, 1)$, where θ_i - cell type proportions for sample *i*.

Then, corresponding numbers of cells were randomly sampled from these three cell types, and the gene expression of these selected cells was then summed to create the final bulk sample. To ensure comparability across samples, we normalized the resulting expression values by the total read count per sample.

For the bulkification of malignant cell states, we used five datasets ([5,21–24]) where state annotations were either directly available or inferred. In cases where annotations were missing, we employed a marker gene scoring approach using Scanpy [40], as described earlier [5], to assign each cell to the state for which it had the highest score. To construct a bulk sample, we aggregated gene expression of cells coming from a particular sample and normalized the expression values by the total read count. For glioblastoma data, where only TPM values were available, we aggregated these values directly without additional normalization.

Comparison with existing methods

We compared CDState performance with four existing methods: CellDistinguisher [25], Linseed [26], Archetype Analysis (AA) [29,30], and debCAM [27,28]. These methods, however, do not provide a dedicated methodology to assign malignant and tumor microenvironment states. To make a fair comparison, we applied our approach to assign these states for all methods (section “Definition of TME and malignant cell states”). Additionally, debCAM, CellDistinguisher and Linseed, perform their own gene filtering, which limits our ability to compare results across all genes. Consequently, the performance comparison was made only using the subset of genes retained by these methods. Furthermore, AA and Linseed do not output proportions that sum to one, therefore we rescaled the proportions from these methods so that they summed to 1.

Bulkified mixtures

For the bulkified mixtures, the methods were run for three components, and each method (except deterministic CellDistinguisher) was run for 20 times. We compared the methods root mean squared error (RMSE) and Pearson’s correlation coefficient with the ground truth proportions and source expression profiles.

Bulkified tumor datasets

Bulkified datasets were generated by aggregating single-cell expression profiles and computing ground truth proportions based on cell-type composition. The cell state annotations were obtained directly from the original studies when available. In cases where annotations were missing, we assigned cell states based on maximal scores obtained using Scanpy [40]. Ground truth gene expression profiles were established by averaging the expression of cells within each identified cell type or state.

To ensure a fair comparison, we ran CDState and the four selected deconvolution methods across a range of component numbers (*k*) ranging from 2 to 15, with the exception of debCAM, which due to the runtime, was not able to be run for k>10 (Additional File 1). Since the existing methods do not inherently distinguish malignant from TME components, we applied the same assignment approach as the one used in CDState. Specifically, for each method and each value of *k*, we selected the components that maximized correlation with the true tumor purity (Additional File 1). This allowed us to determine the number of inferred malignant states per *k*. Knowing the true number of malignant states in each dataset, we selected the *k* that best matched the expected number of malignant states (Suppl. Table S1).

If the exact number was not identified by a method, we chose the solution with the closest higher value.

The performance of each method was evaluated based on the inferred proportions and source gene expression profiles. To quantify the performance, we used three different metrics:

- Pearson's correlation coefficient, defined as $\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum(X_i-\overline{X})^2\sum(Y_i-\overline{Y})^2}}$,,
- Bray-Curtis dissimilarity index, defined as $\frac{\sum|X_i-Y_i|}{\sum(X_i+Y_i)}$,
- and the relative root mean squared error (rRMSE), defined as $\frac{\sqrt{(X_i-Y_i)^2}}{(max(X_i)-min(X_i)}$.

For Pearson's correlation coefficient, we considered only non-negative coefficients to assign inferred states, therefore all the calculated metrics ranged between 0 and 1, allowing direct comparison across datasets and states. Finally, to summarize the overall performance of each method, mean values of the above scores across datasets were calculated and summed, providing an overall performance score for each method.

CDState deconvolution of bulkified glioblastoma data

Bulkified glioblastoma dataset [5] was deconvolved with CDState for 2 to 15 states. We selected the best solution (k=7) based on the highest correlation between estimated and true tumor purity.

Then, to identify marker genes for each inferred cell state, we selected the top 100 genes with the highest expression in a particular source.

Identified marker genes were then used to score the single-cell populations used for bulkification, applying the ANS method [33].

Finally, we followed the approach described in *Neftel et al.* [5] to visualize malignant cells within a two-dimensional space defined by OPC-, NPC-, AC-, and MES-like cell states. The x- and y-axis position of cells was calculated as described in the original study by *Neftel et al.* [5].

Pancancer deconvolution of TCGA data

RSEM-normalized bulk RNA-seq data for all TCGA tumor types were downloaded from Xena Browser (https://xenabrowser.net/datapages/), along with ABSOLUTE tumor purity estimates from the TCGA PanCanAtlas supplemental data (https://gdc.cancer.gov/about-data/publications/pancanatlas). CDState was applied separately to each cancer type, performing deconvolution with *k* ranging from 2 to 15, with 10 independent runs for each value of *k*. The selected number of *k* for each cancer type can be found in Suppl. Table S5.

To determine the optimal number of transcriptional cell states for each cancer type, we used the following selection criteria:

- K which has the highest correlation between estimated and true tumor purity
- At least two malignant cell states identified
- Plateau in the number of malignant states across increasing values of *k*

Calculation of ITTH score

The ITTH score is calculated as follows:

$ITTH = \frac{-\sum_{i=1}^{k} p_i \log p_i}{\log k}$, where p_i represents the proportion of malignant state *i*, and *k* is the total number of malignant cell states. A higher ITTH score indicates increased transcriptional heterogeneity within the tumor.

Gene set enrichment analysis and state annotation

Marker genes of inferred malignant cell states were identified using the procedure described in the section "CDState deconvolution of bulkified glioblastoma data." For each state, a hypergeometric test implemented in the fgsea R package [41] was performed to assess the association between the marker genes and hallmark of cancer gene sets, downloaded from MsigDB (https://www.gsea-msigdb.org/gsea/msigdb). Hallmarks with a p-value < 0.05 were considered significantly enriched. To simplify biological interpretation, the 50 hallmark gene sets were manually grouped into 11 functional categories (Supplementary Table S7). Each malignant cell state was assigned to a specific functional group based on the classification of the hallmark set with the lowest p-value association with the given state (Supplementary Table S8).

Clinical characterization of malignant cell states

The ITTH index and the proportions of identified malignant cell states were analyzed for associations with patient clinical features. The Kruskal-Wallis test was used to assess the relationship between the ITTH score and tumor molecular subtypes, as well as other clinical characteristics of TCGA patients. To evaluate the prognostic relevance of ITTH and malignant cell state proportions, Cox univariate survival analysis was used (using survival and survminer R packages).

Identification of genetic drivers of malignant cell states

Gistic2 copy number data and MC3 gene-level non-silent mutation data for TCGA samples was downloaded from Xena Browser (https://xenabrowser.net/datapages/). For better interpretability, copy-number changes were encoded as binary events for gains (Gistic2 values 1 and 2) or losses (Gistic2 values -1 and -2). For each cancer type, we first filtered out genes with genetic change in less than 10% of samples. Then, we used point biserial correlation for each CDState-identified malignant state proportions and gene alteration profiles, to assess the association. Obtained p-values within each cancer type were adjusted for the number of tests. For analysis of copy number changes of transcriptional factors, we used transcriptional factors defined in the FANTOM5 database (https://fantom.gsc.riken.jp/5/).

Data and code availability

Code availability

CDState is implemented in Python and is freely available at https://github.com/BoevaLab/CDState.

Data availability

The single-cell studies used in this study can be downloaded from:

- Broad Institute Single-Cell Portal: Single cell RNA-seq of adult and pediatric glioblastoma, Neftel et al., https://singlecell.broadinstitute.org/single_cell/study/SCP393/single-cell-rna-seq-of-adult-and-pediatric-glioblastoma; A single-cell and spatially resolved atlas of human

breast cancers, Wu et al., https://singlecell.broadinstitute.org/single_cell/study/SCP1039/a-single-cell-and-spatially-resolved-atlas-of-human-breast-cancers;

- The Gene Expression Omnibus (GEO): Single cell RNA sequencing of lung adenocarcinoma, Kim et al., GSE131907;
- Cellxgene portal: MSK Spectrum - Ovarian cancer mutational processes drive site-specific immune evasion, Vazquez-Garcia et al., https://cellxgene.cziscience.com/collections/4796c91c-9d8f-4692-be43-347b1727f9d8;
- Curated Cancer Cell Atlas: Skin squamous cell carcinoma, Ji et al., https://www.weizmann.ac.il/sites/3CA/skin;

The bulk data used in this study can be downloaded from:

- Xena Browser (https://xenabrowser.net/datapages/): TCGA IlluminaHiSeq (RSEM-normalized) data; Gistic2 thresholded copy number data; MC3 gene-level non-silent mutation data;
- PanCanAtlas supplementary files (https://gdc.cancer.gov/about-data/publications/pancanatlas): TCGA Absolute purity/ploidy file; TCGA-clinical Data Resource Outcome;
- TCGAbiolinks R package (https://rdrr.io/bioc/TCGAbiolinks): full clinical characterization of TCGA samples; subtype annotation of TCGA samples;

Declarations

Ethics approval

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Funding

This work is supported by the Swiss National Science Foundation (SNSF) grant number 205321_207931.

Author information

Authors' contributions

AK developed CDState, performed computational analyses, and prepared the manuscript with input from all authors. AK and VB conceived the study. VB supervised the study. JY prepared single cell data for benchmarking. All authors read and approved the final manuscript.

Acknowledgments

We sincerely thank Prof. Daniela Witten, University of Washington, for inspiring the idea to use purity information within the CDState algorithm.

Supplementary Material

Supplementary Figures

Suppl. Figure S1. Performance comparison between CDState and existing unsupervised deconvolution methods: Archetype Analysis (AA), Linseed, CellDistinguisher, debCAM, and output from the first step of CDState (CW-NMF). Barplots represent Pearson's correlation coefficient and root mean squared error (RMSE) values from comparing inferred and true source gene expression and proportions of B-cells, Malignant cells and Fibroblasts from *Kim. et al.* single cell dataset.

Suppl. Figure S2. Comparison of pairwise cosine similarity between gene expression vectors of B-cells, Fibroblasts and Malignant cells between inferred source gene expression after first step of CDState (CW-NMF) and second step (CDState). Dotted lines represent cosine similarity between ground truth gene expression vectors.

Suppl. Figure S3. Performance comparison in detection of malignant cell states reported in the original studies. Heatmaps represent methods' performance when comparing estimated and true proportions and source gene expression, by evaluating Pearson's correlation coefficient, relative Root Mean Squared Error (rRMSE) and Bray-Curtis dissimilarity index (B-C index). Histograms report the percentage of detected states across the five datasets based on positive correlation with true proportions and source gene expression. / - state was not identified.

Suppl. Figure S4. Selection of optimal number of cell states from CDState deconvolution of bulkified glioblastoma dataset. **A.** Average correlation between CDState-inferred and true tumor purity across 20 CDState runs for a range of *k*. **B.** Number of malignant components identified in the run with the highest correlation between true and inferred tumor purity across a range of *k*.

Suppl. Figure S5. Correlation between true proportions of cells in the bulkified GBM *Neftel et al.* dataset. Numbers correspond to Pearson's correlation coefficient.

Suppl. Figure S6. Correlation between true and CDState-identified cell state proportions of glioblastoma dataset, across a range of *k* (3 to 7). Red font above heatmaps corresponds to malignant cell states.

Suppl. Figure S7. Correlation between true and CDState-identified cell state proportions of glioblastoma dataset, across a range of *k* (8 to 10). Red font above heatmaps corresponds to malignant cell states. .

Suppl. Figure S8. Deconvolution of TCGA datasets with CDState. **A.** Pearson's correlation between CDState-inferred and true tumor purity of TCGA datasets. **B.** Comparison of correlation between gene copy number profile and deconvolved malignant (blue) or bulk (pink) gene expression across TCGA datasets. The deconvolved malignant expression was calculated per sample, by subtracting CDState-inferred tumor-microenvironment source expression multiplied by corresponding inferred proportions from the bulk RNA-seq data.

Suppl. Figure S9. Candidate genetic drivers of TCGA malignant cell heterogeneity. **A.** Gene copy number aberrations in transcriptional factors with the highest and significant (adjusted p-value < 0.05) biserial correlation with cancer-type specific heterogeneity scores. Cancer-specific frequency of event is represented by size of points. Copy number loss and gain is represented by point shape. **B.** Top 30 genes with copy number aberrations most

frequently associated with TCGA malignant cell states (adjusted p-value < 0.05). Numbers on top of bar plots correspond to the mean frequency [%] of events (mutation or copy number change) across all analyzed TCGA samples. Blue bars represent gene copy number loss, pink bars represent gene copy number gain.

Suppl. Figure S10. Gene copy number changes in transcriptional factors with the highest significant association (adjusted p-value < 0.05) with malignant state proportions which are recurrently linked to states across cancer types. Cancer-specific frequency of event is represented by size of points. Copy number loss and gain is represented by point shape.

Supplementary Tables

Supplementary Table S1. Number of components selected for the benchmark on bulkified tumor datasets.

Supplementary Table S2. Summary statistics of proportion inference across bulkified cancer datasets and methods.

Supplementary Table S3. Summary statistics of source gene expression inference across bulkified cancer datasets and methods.

Supplementary Table S4. Overall performance of CDState and existing methods across bulkified cancer datasets.

Supplementary Table S5. Identified optimal number of components across TCGA cancer types.

Supplementary Table S6. Results of Cox univariate survival analysis across malignant cell states and ITTH index identified for TCGA cancers.

Supplementary Table S7. Functional grouping of MsigDB hallmarks of cancer sets.

Supplementary Table S8. Results of gene set enrichment analysis using MsigDB hallmarks of cancer, on the CDState-inferred malignant cell states across TCGA cancers.

Supplementary Table S9. Recurrent hallmark groups expressed in malignant cell states across TCGA cancer types.

Supplementary Table S10. Number of malignant cell states across values of k and methods used for benchmarking.

Additional Data

Additional File 1.pdf

Additional methods and results used to select the optimal number of states across analyzed dataset

References

1. Easwaran H, Tsai H-C, Baylin SB. Cancer epigenetics: tumor heterogeneity, plasticity of stem-like states, and drug resistance. Mol Cell. 2014;54: 716–727.

2. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell. 2017;168: 613–628.

3. Alizadeh AA, Aranda V, Bardelli A, Blanpain C, Bock C, Borowski C, et al. Toward understanding and exploiting tumor heterogeneity. Nat Med. 2015;21: 846–853.

4. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344: 1396–1401.

5. Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, et al. An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell. 2019;178: 835–849.e21.

6. Baron M, Tagore M, Hunter MV, Kim IS, Moncada R, Yan Y, et al. The Stress-Like Cancer Cell State Is a Consistent Component of Tumorigenesis. Cell Syst. 2020;11: 536–546.e7.

7. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell. 2017;171: 1611–1624.e24.

8. Kim C, Gao R, Sei E, Brandt R, Hartman J, Hatschek T, et al. Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. Cell. 2018;173: 879–893.e13.

9. Izar B, Tirosh I, Stover EH, Wakiro I, Cuoco MS, Alter I, et al. A single-cell landscape of high-grade serous ovarian cancer. Nat Med. 2020;26: 1271–1279.

10. Rambow F, Rogiers A, Marin-Bejar O, Aibar S, Femel J, Dewaele M, et al. Toward Minimal Residual Disease-Directed Therapy in Melanoma. Cell. 2018;174: 843–855.e19.

11. Crawford Y, Kasman I, Yu L, Zhong C, Wu X, Modrusan Z, et al. PDGF-C mediates the angiogenic and tumorigenic properties of fibroblasts associated with tumors refractory to anti-VEGF treatment. Cancer Cell. 2009;15: 21–34.

12. Dirkse A, Golebiewska A, Buder T, Nazarov PV, Muller A, Poovathingal S, et al. Stem cell-associated heterogeneity in Glioblastoma results from intrinsic tumor plasticity shaped by the microenvironment. Nat Commun. 2019;10: 1787.

13. Cazet AS, Hui MN, Elsworth BL, Wu SZ, Roden D, Chan C-L, et al. Targeting stromal remodeling and cancer stem cell plasticity overcomes chemoresistance in triple negative

breast cancer. Nat Commun. 2018;9: 2897.

14. Barkley D, Moncada R, Pour M, Liberman DA, Dryg I, Werba G, et al. Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. Nat Genet. 2022;54: 1192–1201.

15. Gavish A, Tyler M, Greenwald AC, Hoefflin R, Simkin D, Tschernichovsky R, et al. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. Nature. 2023;618: 598–606.

16. Jew B, Alvarez M, Rahmani E, Miao Z, Ko A, Garske KM, et al. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nat Commun. 2020;11: 1971.

17. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;37: 773–782.

18. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019;10: 380.

19. Dong M, Thennavan A, Urrutia E, Li Y, Perou CM, Zou F, et al. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. Brief Bioinform. 2021;22: 416–427.

20. Chu T, Wang Z, Pe'er D, Danko CG. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. Nat Cancer. 2022;3: 505–517.

21. Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun. 2020;11: 1–15.

22. Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, et al. A single-cell and spatially resolved atlas of human breast cancers. Nat Genet. 2021;53: 1334–1347.

23. Vázquez-García I, Uhlitz F, Ceglia N, Lim JLP, Wu M, Mohibullah N, et al. Ovarian cancer mutational processes drive site-specific immune evasion. Nature. 2022;612: 778–786.

24. Ji AL, Rubin AJ, Thrane K, Jiang S, Reynolds DL, Meyers RM, et al. Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. Cell. 2020;182: 497–514.e22.

25. Newberg LA, Chen X, Kodira CD, Zavodszky MI. Computational de novo discovery of distinguishing genes for biological processes and cell types in complex tissues. PLoS One. 2018;13: e0193067.

26. Zaitsev K, Bambouskova M, Swain A, Artyomov MN. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. Nat Commun. 2019;10: 2209.

27. Chen L, Wu C-T, Wang N, Herrington DM, Clarke R, Wang Y. debCAM: a bioconductor R package for fully unsupervised deconvolution of complex tissues. Bioinformatics. 2020;36: 3927–3929.

28. Wang N, Hoffman EP, Chen L, Chen L, Zhang Z, Liu C, et al. Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex

tissues. Sci Rep. 2016;6: 18909.

29. Eugster MJA, Leisch F. From spider-man to hero - archetypal analysis inR. J Stat Softw. 2009;30. doi:10.18637/jss.v030.i08

30. Eugster MJA, Leisch F. Weighted and robust archetypal analysis. Comput Stat Data Anal. 2011;55: 1215–1225.

31. Nguyen H, Nguyen H, Tran D, Draghici S, Nguyen T. Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges. Nucleic Acids Res. 2024;52: 4761–4783.

32. Groves SM, Ildefonso GV, McAtee CO, Ozawa PMM, Ireland AS, Stauffer PE, et al. Archetype tasks link intratumoral heterogeneity to plasticity and cancer hallmarks in small cell lung cancer. Cell Syst. 2022;13: 690–710.e17.

33. Ciernik L, Kraft A, Barkmann F, Yates J, Boeva V. ANS: Adjusted Neighborhood Scoring to improve gene signature-based cell annotation in single-cell RNA-seq data. bioRxiv. 2023. doi:10.1101/2023.09.20.558114

34. Wang S, Chen F-E. Small-molecule MDM2 inhibitors in clinical trials for cancer therapy. Eur J Med Chem. 2022;236: 114334.

35. Sabapathy K, Lane DP. Therapeutic targeting of p53: all mutants are equal, but some mutants are more equal than others. Nat Rev Clin Oncol. 2018;15: 13–30.

36. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun. 2015;6: 8971.

37. Haider S, Tyekucheva S, Prandi D, Fox NS, Ahn J, Xu AW, et al. Systematic Assessment of Tumor Purity and Its Clinical Implications. JCO Precis Oncol. 2020;4. doi:10.1200/PO.20.00016

38. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30: 413–421.

39. Kirkpatrick S, Gelatt CD Jr, Vecchi MP. Optimization by simulated annealing. Science. 1983;220: 671–680.

40. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19: 15.

41. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. bioRxiv. bioRxiv; 2016. doi:10.1101/060012