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], 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 the 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 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 reference datasets, limiting the performance of such methods by the quality of used reference data. On the other hand, unsupervised approaches do not face such limitations, but existing methods have not been specifically adapted to characterize malignant cell states, focusing 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 (NMF) 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 the TCGA, we identified recurrent malignant cell programs, including epithelial-mesenchymal transition (EMT) and hypoxia as main drivers of tumor transcriptional heterogeneity. We further link the identified malignant states to patient clinical features, revealing states associated with worse patient prognosis. Finally, we find that alterations in genes such as *KRAS* or *TP53*, whose copy number or mutation status strongly correlate with specific malignant states, may play a key role in driving these cellular states. Overall, CDState provides a powerful and accessible approach to resolving intratumor heterogeneity using bulk RNA-seq data and emerges as a promising tool for advancing our understanding of malignant cell heterogeneity.

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 most of the 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 [4,10–12]. These states are closely intertwined with TME composition, influencing processes such as angiogenesis [13], and immune evasion [7,14,15]. 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 [16,17]. 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 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 [18–22]. However, most of these methods rely on reference transcriptomic data cell populations within the bulk data, 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 single-cell reference, limiting their ability to accurately resolve tumor-specific transcriptional heterogeneity.

To address these limitations, we introduce CDState, a novel 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 [23], breast cancer [24], ovarian high grade serous carcinoma [25], and squamous cell carcinoma [26], showing that it outperforms existing unsupervised deconvolution methods [27–32] in both cell state proportions and gene expression estimates across diverse tumor contexts.

Applying CDState to 33 cancer types from The Cancer Genome Atlas (TCGA), we uncover recurrent gene expression programs driving ITTH across cancers, with epithelial-mesenchymal transition (EMT), hypoxia, and immune response, emerging as dominant transcriptional drivers. We further establish strong associations between malignant

cell states and clinical outcomes, identifying states linked to more aggressive disease and worse prognosis in multiple cancers. In addition, we show that specific genetic alterations, such as *KRAS* or *TP53* mutations and copy number changes, are significantly associated with particular malignant states, suggesting potential molecular mechanisms underlying tumor heterogeneity.

By offering an accessible and scalable solution for dissecting intratumor heterogeneity from bulk RNA-seq, CDState offers a promising solution to expand our understanding of intratumor heterogeneity of tumor, potentially highlight new molecular targets for future treatments and patient stratification.

Results

Overview of CDState

CDState is a novel method designed to infer malignant cell states and their proportions from bulk RNA-sequencing data (Figure 1A). CDState is a framework based on Nonnegative Matrix Factorization (NMF) with additional constraints to improve the accuracy and interpretability of deconvolution.

Compared to other NMF-based methods that rescale weights after deconvolution, CDState imposes sum-to-one constraints on estimated weights into the optimization process (Figure 1B, Methods). Furthermore, to improve the separation of malignant cell states, we introduce a novel objective function term that penalizes high cosine similarity between deconvolved source gene expression profiles. By reducing similarity, CDState effectively distinguishes transcriptionally similar cell states, such as 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). In the first step, CDState prioritizes minimizing reconstruction error ($\beta = 0$). In the second step, β is gradually increased, giving more importance to separation of source gene expression profiles. While temporarily allowing for higher reconstruction error, this approach aims to improve source separation. 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. Once the MSE stops decreasing, β is gradually reduced back to 0 for a final optimization step that minimizes reconstruction error (Figure 1B, Methods).

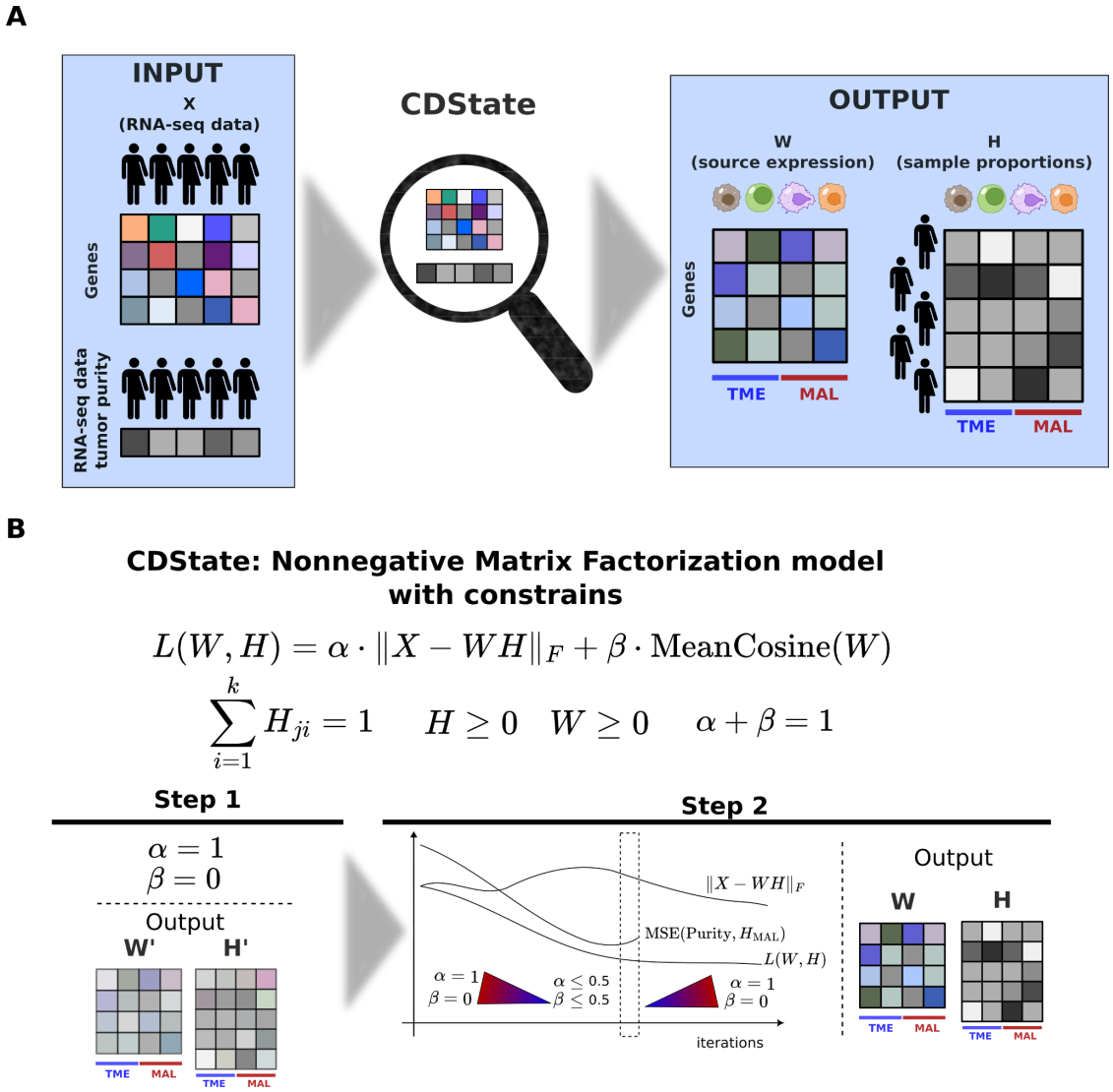

Figure 1. CDState overview. **A.** CDState takes as input a normalized RNA-seq expression matrix along with sample purity estimates and returns in output inferred proportion and source cell state expression matrix, along with assignment to tumor microenvironment (TME) or malignant (MAL) cell states. **B.** CDState algorithmic flow and parameters.

CDState offers the best performance for cell type proportion and gene expression deconvolution compared to existing methods

We first tested CDState on bulkified single-cell RNA sequencing data from lung adenocarcinoma (*Kim et al.* [23]). 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) (Figure 2B).

We compared CDState against four state-of-the-art unsupervised deconvolution methods: Archetype Analysis (AA) [31,32], debCAM [29,30], Linseed [28], and CellDistinguisher [27]. These methods were selected based on a recent benchmark highlighting them as top-performing deconvolution tools [33], while AA was included as it was previously used to study molecular phenotypes in cancer [34]. 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 RMSE with both ground truth source profiles and proportions (Figure 2B, Suppl. Figure S2). Notably, it showed the smallest variability in performance metrics across multiple runs, indicating robust and stable results. Importantly, CDState performed the best in the particularly challenging MixD scenario, which closely reflects bulk RNA-seq sample composition (Suppl. Figure S2). Furthermore, the cosine similarity-based optimization improved CDState performance, by reducing RMSE and improving correlation with true source profiles (Suppl. Figure S2) compared to the CW-NMF results. 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 S3).

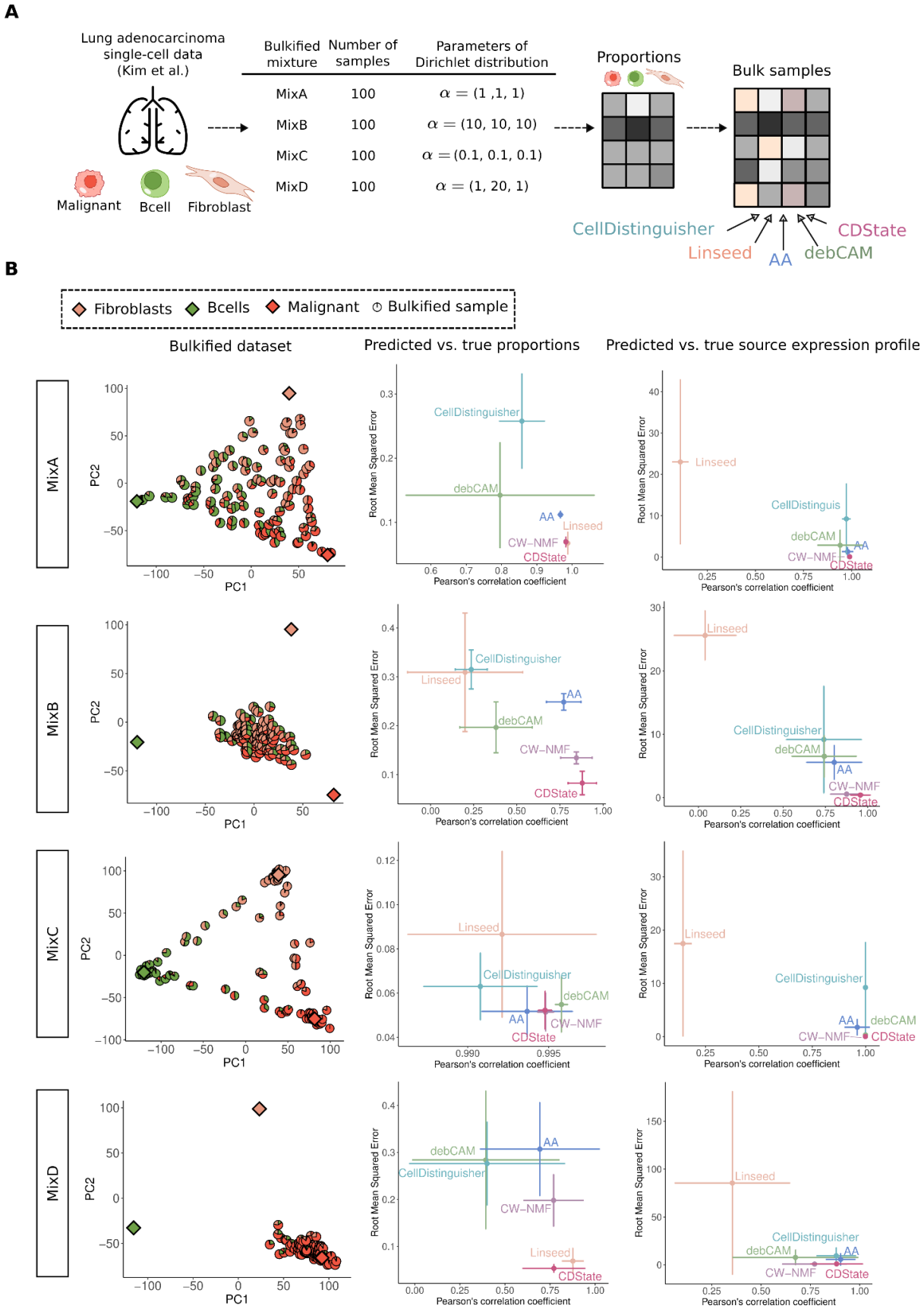

Figure 2. Comparison of CDState performance with other deconvolution methods using simulated data with varying degree of mixing proportions. **A.** Single-cell lung adenocarcinoma data from Kim et al. was used to create four mixtures for benchmarking. Dirichlet distribution with varying alpha was

used to draw proportions of three cell types (malignant cells, B cells and fibroblasts). Given the sampled proportions, corresponding numbers of cells were selected and their gene expression was summed to create bulkified samples. CDState along with 4 other methods was used to validate the deconvolution performance of these mixtures. **B.** Deconvolution performance on four bulkified mixtures (rows) using four external methods as well as output from the first step of CDState (CW-NMF). 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 in terms of Root Mean Squared Error and Pearson's correlation coefficient on both inferred proportions (middle column) and source gene expression profiles (right column).

CDState shows better performance in gene expression and proportion inference of malignant cell states compared to other methods

To further evaluate CDState's performance in deconvolving malignant cell states, we applied it to five bulkified single-cell RNA-seq datasets from breast [24], ovarian [25], glioblastoma [5], lung [23], and squamous cell carcinoma [26] cancer datasets (Figure 3A).

We first assessed each method's ability to distinguish malignant cells from tumor microenvironment (TME) components. CDState consistently outperformed other methods, achieving the highest correlation with tumor purity and the lowest RMSE in all datasets (Figure 3B). Then, we verified CDState's performance in inferring malignant cell states and their proportions. For each malignant state, we computed Pearson's correlation coefficient, relative RMSE (rRMSE), and Bray-Curtis dissimilarity index (Figure 3C, Methods) between the true and inferred source gene expression profiles and proportions.

CDState showed better performance across multiple metrics compared to existing deconvolution methods (Figure 3C, Suppl. Table S4). It achieved the highest correlation with gene expression sources (R = 0.797) (Suppl. Table S4), and the second highest correlation with proportions (R=0.513). While it achieved worse rRMSE and Bray-Curtis index for proportions (0.460) compared to other methods, it showed a better performance in recovering gene expression of sources.

Additionally, CDState ranked second in identifying malignant states based on correlation with proportions (28/30 positively correlated states) and identified all malignant cell states when evaluating source gene expression (30/30 positively correlated states) (Figure 3C).

While some ground truth states were not captured by separate CDState-inferred states, these states could still be detected in some of the deconvolved sources. For instance, inferred proportions of two states: GBM state G2.M and BRCA state mod6, were highly correlated with the proportions of already assigned states (Suppl. Figure S6). This could be caused by selecting a CDState solution with too few states, making it more challenging for the method to separate biologically similar states.

However, across all datasets, CDState achieved the best overall performance (Figure 3D, Suppl. Table S2-S4).

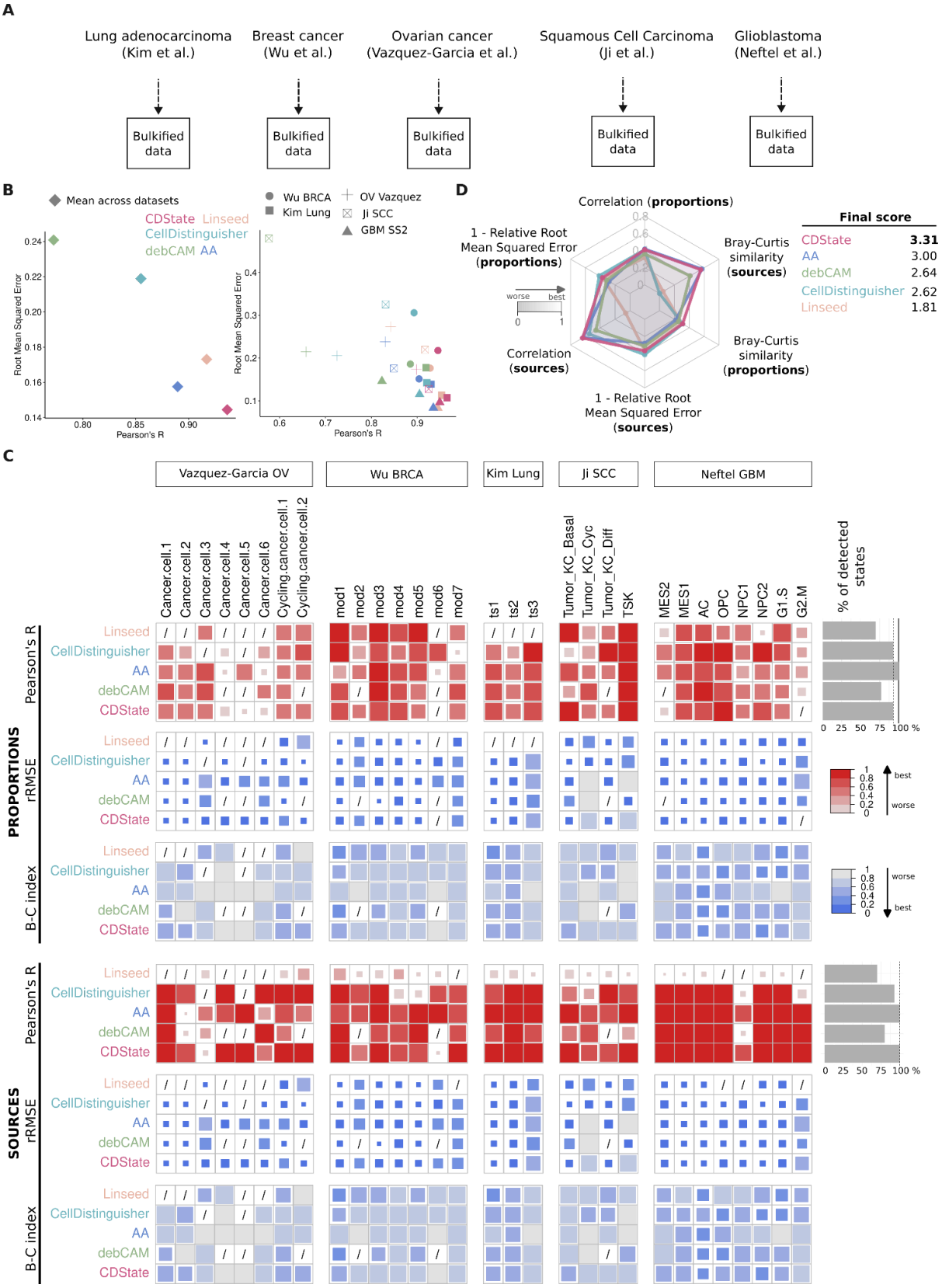

Figure 3. CDState performance validation using five bulkified tumor single cell datasets. **A.** Datasets used for benchmark. For each dataset separately, cells were aggregated per patient and corresponding cell proportions were used as ground truth. **B.** Comparison of deconvolved and true tumor purity across the five datasets (left) and per dataset (right). Point shapes correspond to the

dataset, while colors mark the deconvolution method. **C.** 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. **D.** Performance summary across the five 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 accurately infers gene expression and proportions of glioblastoma malignant cell states

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

We determined the optimal number of components by selecting the number of states (k) that maximized the mean correlation with tumor purity across 20 runs. This approach led to selecting a seven-state solution (Suppl. Figure S7). After identifying malignant and TME components, we evaluated their correlation with the ground truth proportions of corresponding cell types (Figure 4A). CDState showed high correlation between true and inferred proportions, showing clear separation of MES1-, NPC1-, and OPC-like states, with AC- and MES2-like, as well as NPC2- and NPC1-like, were expressed together (Figure 4A-B).

Further, tumor microenvironment cells, including macrophages and T-cells, were correctly separated from malignant cells, however, CDState failed to separate a source corresponding to oligodendrocytes, likely due to their low proportion in the dataset.

To further validate these findings, we scored single cells in the bulkified dataset using marker genes derived from CDState's inferred states. The resulting scores aligned well with the true cell state assignments (Figure 4B). Additionally, we mapped these single-cell scores onto the transcriptional space of the original *Neftel et al.* dataset. CDState successfully identified the main axes of variation, with two OPC-NPC gradients (one stronger for OPC-like cells, the other for NPC-like cells) and two AC-MES gradients (one stronger for AC-like cells, the other for MES-like cells) (Figure 4C).

Importantly, CDState also identified cycling cell populations (states S1, S3, and S6), consistent with the findings of *Neftel et al.* [5]. These results support the observation that increasing *k* allows for finer resolution of state separation (Figure 4C). Our analysis of state scores across increasing values of *k* (Suppl. Figure S8), further confirmed CDState's ability to separate malignant cell states into distinct subpopulations.

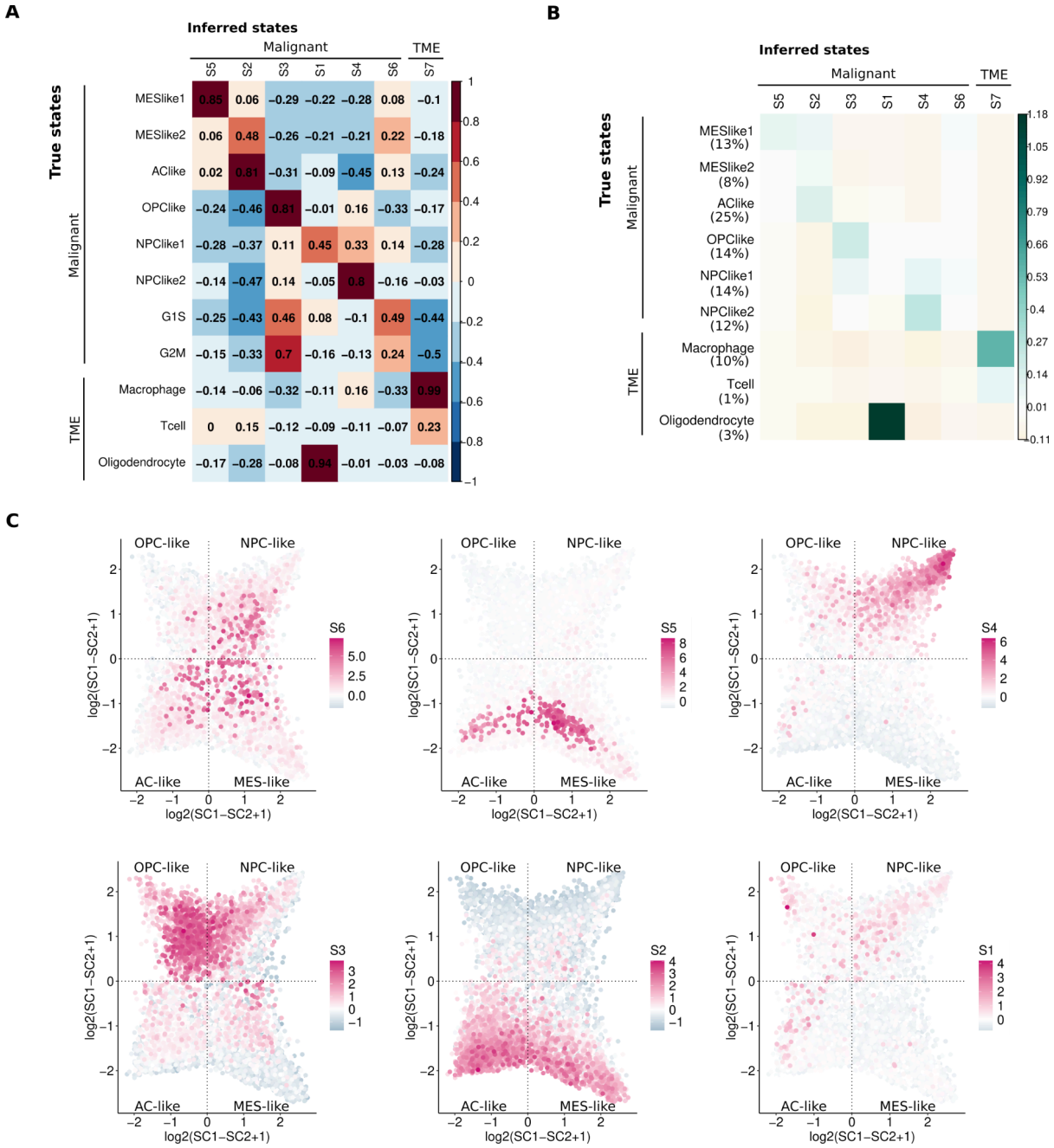

Figure 4. CDState deconvolves cell states in bulkified glioblastoma data. **A.** Heatmap with Pearson's correlation coefficient between true and CDState estimated proportions. **B.** Single cell scores of CDState deconvolved states. Scores were calculated using ANS method (*Ciernik et al.*). Percentages represent the proportion of cells in the bulkified dataset. **C.** Two-dimensional representation of glioblastoma malignant cell states from Neftel et al. 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

We applied CDState to The Cancer Genome Atlas (TCGA) datasets, to provide a pan-cancer characterization of malignant cell state heterogeneity. Specifically, we analyzed bulk RNA-seq data from 33 cancer types, across 9,669 patient samples (Figure 5A).

Overall, CDState achieved good separation between malignant and TME components - 24 out of 33 TCGA cancer datasets achieved correlation of at least 0.5 between estimated and true tumor purity (Suppl. Figure S10A). Additionally, in 32 of 33 cancer types, deconvolved malignant gene expression profiles showed stronger association with copy number alterations compared to bulk RNA-seq data (Suppl. Figure S10B).

Analysis of malignant cell proportions revealed variability in transcriptional heterogeneity across TCGA cancer types, with stomach adenocarcinoma (STAD) showing the lowest ITTH values, and adrenocortical carcinoma (ACC) the highest (Figure 5B).

To characterize the molecular activity of the inferred malignant states, we examined their association with MSigDB hallmarks of cancer. Many states expressed recurrent molecular programs across cancer types, with strong enrichment for Epithelial-mesenchymal transition (EMT) (27/33 cancer types) – a process linked to tumor development and differentiation; Hypoxia (19/33 cancer types) – process associated with cellular stress and adaptation, and Coagulation (18/33 cancer types) – related to immune response and inflammation (Suppl. Table S7). These findings suggest that while each cancer type harbors distinct malignant states, certain core transcriptional programs are conserved across malignancies (Figure 5D, Suppl. Table S7-S10).

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 5C, Suppl. Table S6, p-value < 0.05). Interestingly, in clear cell renal carcinoma (KIRC), rectum adenocarcinoma (READ), stomach adenocarcinoma (STAD), and head and neck squamous cell carcinoma (HNSC), higher transcriptional heterogeneity correlated with worse prognosis. Conversely, in ovarian cancer (OV), higher heterogeneity was linked to a lower hazard.

At the malignant cell state level, malignant states in 15 cancer types were significantly associated with worse prognosis, while states in 12 cancer types were linked to better survival (Figure 5E, 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 inflammation. In contrast, states linked to better prognosis were primarily involved in proliferation and immune-related pathways (Suppl. Table S7, S8).

Further, some of the malignant cell states were significantly associated with molecular subtypes and patient clinical features (Figure 5F-H). We observed the strongest correlation in breast cancer (BRCA), where distinct molecular subtypes showed higher proportions of specific malignant states. Specifically, we found higher proportions of state S5 (linked to apical junction pathways) in Basal-like tumors; state S2 (associated with metabolic processed) in HER2-positive tumors; state S7 (linked to development and differentiation) in

Luminal A and Luminal B (LumA, LumB) tumors, and a higher proportion of both S7 and S5 states in Normal-like tumors presented (Figure 5H). Additionally, several states showed significant associations with tumor stage, tissue or organ of origin, and morphology across nine cancer types (Figure 5G).

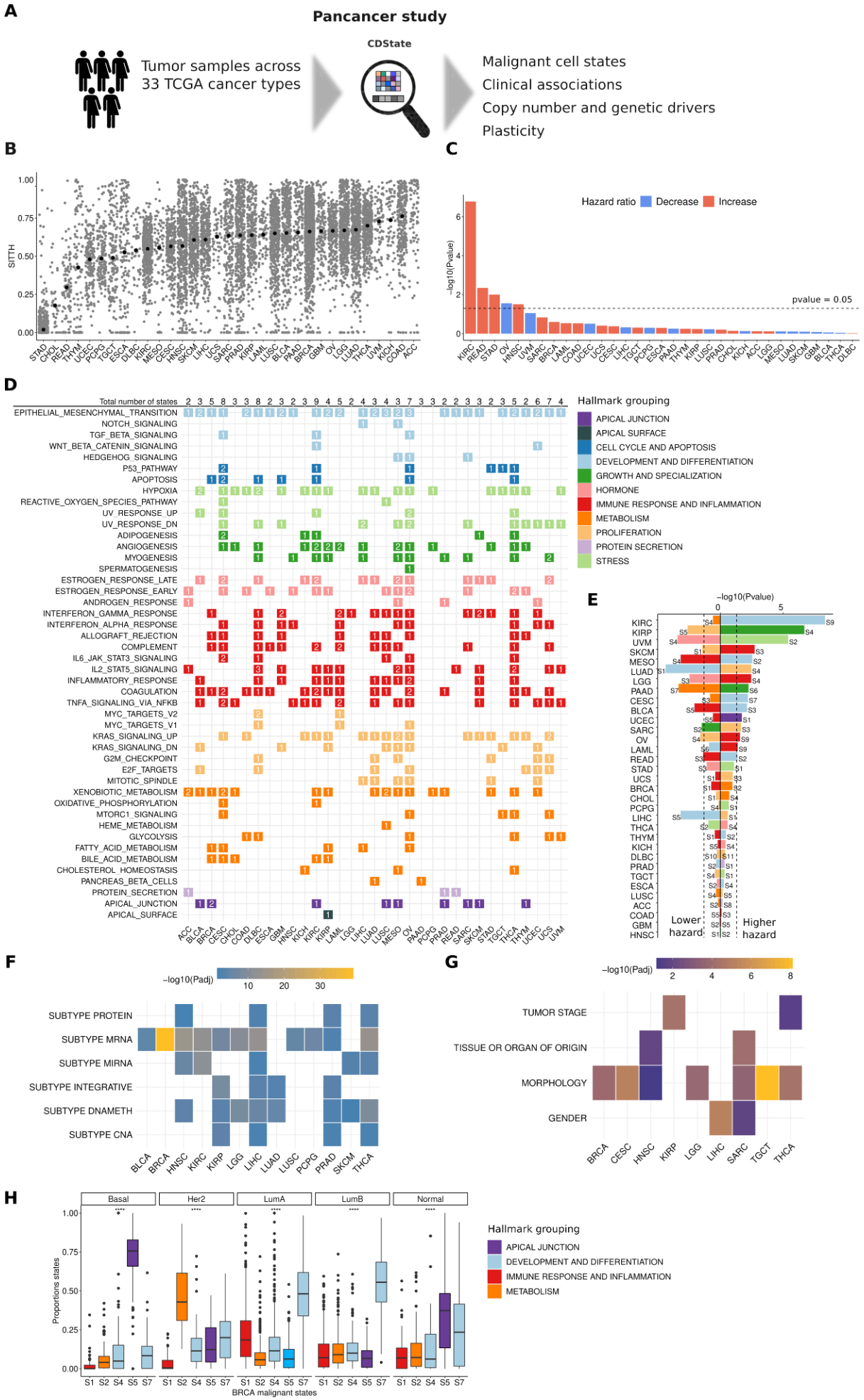
A
Pancancer study
Tumor samples across 33 TCGA cancer types
CDState
Malignant cell states
Clinical associations
Copy number and genetic drivers
Plasticity
B
SITTH
C
Hazard ratio
Decrease
Increase
−log10(Pvalue)
pvalue = 0.05
D
Total number of states
Hallmark grouping
APICAL JUNCTION
APICAL SURFACE
CELL CYCLE AND APOPTOSIS
DEVELOPMENT AND DIFFERENTIATION
GROWTH AND SPECIALIZATION
HORMONE
IMMUNE RESPONSE AND INFLAMMATION
METABOLISM
PROLIFERATION
PROTEIN SECRETION
STRESS
E
−log10(Pvalue)
Lower hazard
Higher hazard
F
−log10(Padj)
SUBTYPE PROTEIN
SUBTYPE MRNA
SUBTYPE MIRNA
SUBTYPE INTEGRATIVE
SUBTYPE DNAMETH
SUBTYPE CNA
G
−log10(Padj)
TUMOR STAGE
TISSUE OR ORGAN OF ORIGIN
MORPHOLOGY
GENDER
H
Basal
Her2
LumA
LumB
Normal
Proportions states
BRCA malignant states
Hallmark grouping
APICAL JUNCTION
DEVELOPMENT AND DIFFERENTIATION
IMMUNE RESPONSE AND INFLAMMATION
METABOLISM

Figure 5. Pancancer characterization of malignant cell heterogeneity using CDState. **A.** RNA-seq data from 33 cancer types deposited in The Cancer Genome Atlas (TCGA) was deconvolved with CDState. The identified malignant cell states have been used to evaluate their clinical significance, i.e. characterize their molecular profiles with enriched hallmarks of cancer, evaluate their link with patient clinical features, and identify potential genetic drivers. **B.** TCGA cancer types present a varied degree of intratumor transcriptional heterogeneity (ITTH), calculated as Shannon entropy of deconvolved proportions. **C.** Association between ITTH and patient survival across TCGA datasets. P-value was calculated using Cox univariate regression analysis. **D.** 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. **E.** 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. **F-G.** Heatmaps show significant (Kruskal-Wallis p-value < 0.05) associations between ITTH and tumor molecular subtypes (**F**) and patient clinical characteristics (**G**). **H.** Distribution of inferred malignant state proportions across molecular subtypes of TCGA BRCA tumors. Colors correspond to the molecular group of the most enriched hallmark in a particular state. Stars represent one-way ANOVA p-value. ns: $p > 0.05$; *: $0.01 < p \leq 0.05$; **: $0.001 < p \leq 0.01$; ***: $0.0001 < p \leq 0.001$; ****: $p \leq 0.0001$.

Genetic drivers of intratumor transcriptional plasticity

To identify candidate genes whose mutation and copy number aberration profiles could promote expression of specific malignant cell states, we used elastic net regression to select genes with the highest contribution in malignant state proportion prediction. Across the 33 cancer types, gene mutations were the most prevalent events (63 states), followed by gene copy number gains (29 states) and losses (25 states) (Figure 6A). Interestingly, some of the mutations with the strongest contribution were associated with several states. For instance, mutations in *BRAF* (THCA states S3 and S5), *GATA3* (BRCA S7 and S1), and *NRAS* (THCA S2 and S4) had the highest contribution to proportions of states within the same cancer type, while several others, including *CTNNB* (LIHC S2 and UCEC S5), *EYS* (ACC S2 and DLBC S4), *KRAS* (UCEC S1, UCS S6), *NFE2L2* (BLCA S3, CESC S2, LUSC S4), *RB1* (BLCA S5, CESC S6, SARC S2), and *SETD2* (KIRP S2 and SKCM S4), were recurrently linked with states across cancer types, (Figure 6A, Suppl. Table S11).

Moreover, among all significant genes, several transcriptional factors were identified as the top affected genes for both mutations (i.e. *TP53*, *NFE2L2*, *ARID1B*, *CREBBP*, Figure 6B) and copy number changes (*E2F3*, *NFIB*, and *NPAS3*, Figure 6C). Particularly interesting are top mutated genes, including *TP53*, *PIK3CA*, *KRAS*, *NFE2L2* and *PTEN*, which has been previously identified as key genetic drivers and potential treatment targets across multiple cancer types [35].

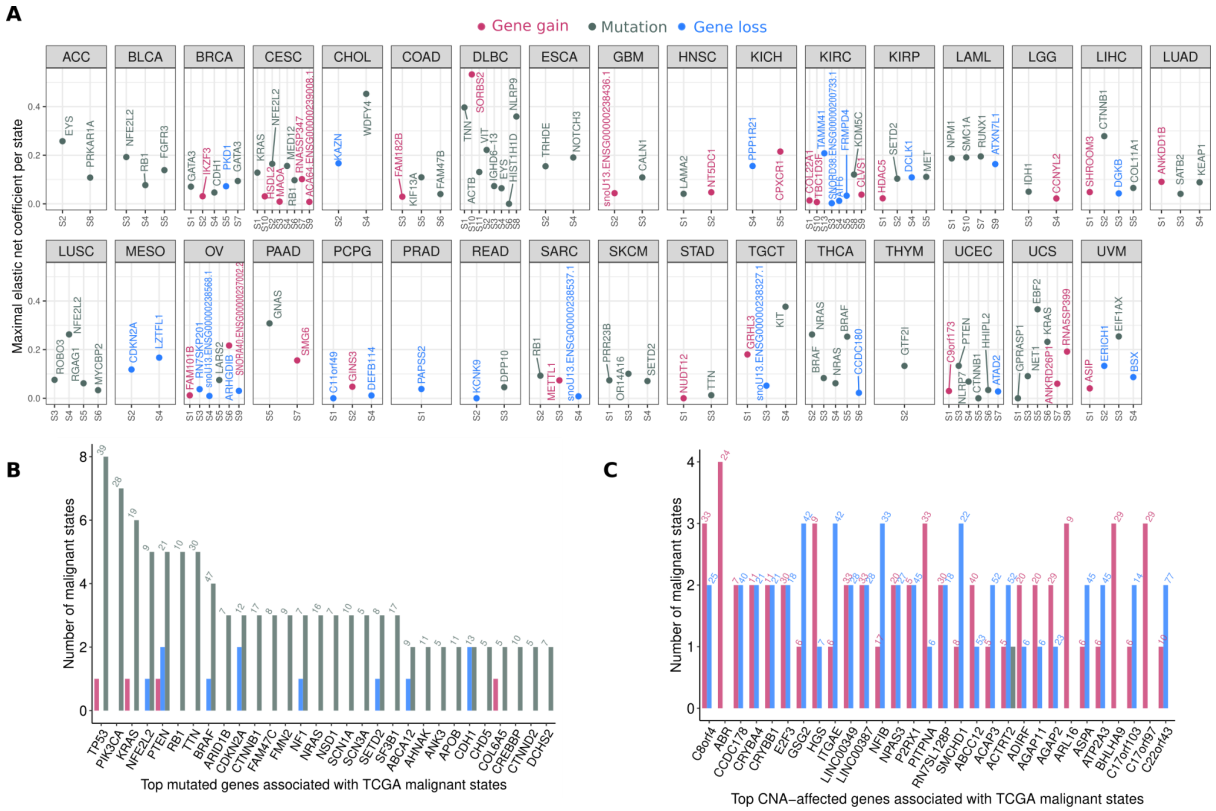

Figure 6. Candidate genetic drivers of TCGA malignant cell states. **A.** Gene mutations and copy number aberrations with the highest elastic net coefficients per TCGA malignant cell states. **B-C.** Top 30 mutated genes (**B**) and top 30 genes with copy number aberrations (**C**) most frequently associated with TCGA malignant cell states. Numbers on top of bar plots correspond to the mean frequency [%] of events (mutation or copy number change) across all TCGA samples.

Discussion

Multiple published studies provide evidence of 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, leaving our understanding limited for many cancer types.

At the same time, large cancer atlases, such as The Cancer Genome Atlas, which collect bulk RNA-seq tumor datasets profiled across multiple patients and cancer types, represent a promising resource for exploring intratumor transcriptional heterogeneity. In this work, we developed CDState, a method designed to analyze such datasets to characterize malignant cell states and expand our understanding of their transcriptional diversity.

We tested CDState across several bulkified datasets and showed its performance advantage compared to other existing unsupervised deconvolution approaches across. Furthermore, we applied CDState to 33 cancer types from TCGA, providing a

comprehensive characterization of transcriptional heterogeneity in malignant cell states at a pan-cancer level. Our analysis identified recurrent malignant cell programs, with EMT and hypoxia emerging as key drivers of malignant cell heterogeneity. Several of the identified states showed strong associations with molecular subtypes and patient clinical characteristics, including patient survival, emphasizing their relevance in patient stratification and potentially treatment selection in future.

Additionally, we identified genetic alterations in known oncogenes as potential drivers of specific malignant cell states, both within and across cancer types. These findings may be particularly important for the development of novel diagnostic approaches and selection of new therapeutic targets, especially for cancer types with low survival rates.

CDState is a promising alternative to supervised deconvolution approaches when accurate single-cell references are unavailable. The main advantage of CDState lies in its ability to accurately deconvolve gene expression profiles of malignant cell states, which is crucial for correctly characterizing biological functions and molecular processes expressed in these cells.

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 (cNMF) with an iterative optimization strategy inspired by simulated annealing [36], to maximize the dissimilarity between gene expression profiles of inferred transcriptional states.

Step 1: Identification of initial solution

1) CDState applies cNMF 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, focusing purely on reconstruction error minimization, runs until convergence.

Constrained Nonnegative Matrix Factorization (cNMF)

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

$x_i = \sum_{k=1}^{K} w_{ik} h_k$ given K sources, where x_i is bulk gene expression of gene i, s_{ik} is gene expression of gene i in pure source k, and w_k is a proportion of source k in a bulk sample. 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:

$\sum_{i=1}^{k} h_{ji} = 1$ for each sample j.

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

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

$\sum_{i=1}^{k} h_{ji} = 1$, $H, W \geq 0$, $\alpha + \beta = 1$, 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 ($\alpha = 1$, $\beta = 0$). 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 [30]. 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 (Suppl. Figure S1). 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 correlation with 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 best number of cell states

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 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.* lung adenocarcinoma dataset [23]: 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],[23],[24],[25],[26]) where state annotations were either directly available or inferred. In cases where annotations were missing, we employed a marker gene scoring approach using Scanpy [37], as described before [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 [27], Linseed [28], Archetype Analysis (AA) [31,32], and debCAM [29,30]. 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 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 [37]. 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 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 (Suppl. Figure S4). 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 (Suppl. Figure S4 - S5). 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 used the following procedure:

a. For each cell state k (column in matrix W), we selected genes with expression values above the 75th percentile within that state
b. The state-specific expression of these genes was compared to the maximum expression of these genes across all other states, using the formula:

$$\frac{W_s[gene,k]-max(expression\ of\ gene\ in\ other\ states)}{max(expression\ of\ gene\ in\ other\ states)}$$

c. The selected genes were ranked in descending order based on this comparison
d. The top 100 genes with the highest relative expression difference were selected as state-specific markers.

Identified marker genes were then used to score the single-cell populations used for bulkification, applying the ANS method [38], which provides a more comparable range of scores for cell state annotation.

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 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 20, with 20 independent runs for each value of *k*.

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 [39], 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). Additionally, one-way ANOVA was used to compare the distribution of malignant cell state proportions across TCGA-BRCA molecular subtypes.

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 each CDState-identified malignant state, elastic net regression was run to identify genes with genetic changes that contribute to proportions of a particular state. First, mutation and copy number matrices were one-hot encoded, with "1" identifying gene mutations, gene copy number loss, or gain. Then, genes were filtered to include only those affected in at least 5% of samples in a given cancer type. Further, genes were tested for association with analyzed state proportions and only genes with significant association were kept (Wilcoxon's test p-value < 0.05). Finally, elastic net regression (implemented in glmnet R package) was run with 10-fold cross validation-based selection of optimal alpha (ranging from 0.5 to 1) and lambda parameters.

Data and code availability

Code availability

CDState is implemented in Python and will be freely available at https://github.com/BoevaLab/, along with the code used to analyze the data and arrive at the conclusions of the study.

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 performed computational analyses, developed CDState 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. Overall magnitude of expression and standard deviation of all and filtered genes for deconvolution with CDState. Marker genes of B-cells, Fibroblasts and Malignant cells from *Kim et al.* lung adenocarcinoma single-cell data are colored with blue, pink and red, respectively. For the three cell types, 50 top marker genes were selected (Methods) using true gene expression of these cell types. The numbers in tables report the number of marker genes kept after applying CDState gene filtering procedure before deconvolution.

Suppl. Figure S2. 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 S3. 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 line represent cosine similarity between ground truth gene expression vectors.

Suppl. Figure S4. Correlation between estimated and true purity along with number of detected malignant cell states across a range of *k* and bulkified single-cell cancer datasets, inferred by existing deconvolution methods.

Suppl. Figure S5. Average correlation and average mean squared error (MSE) between estimated and true purity along with number of detected malignant cell states across 20 runs for a range of *k* and bulkified single-cell cancer datasets, inferred by CDState. Error bars represent the smallest and the highest number of detected malignant cell states for a given *k*.

Suppl. Figure S6. Pearson's correlation coefficients between inferred and true proportions of malignant cell states, obtained by CDState and the four existing unsupervised deconvolution methods, across five bulkified single-cell cancer datasets. Black border represents assignment of inferred cell states (rows) to the ground truth labels (columns).

Suppl. Figure S7. 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 S8. Single cell scores of CDState-identified cell sources, across cell types of glioblastoma dataset and a range of *k*. For each *k*, marker genes were selected using the inferred source gene expression of the run with the highest correlation between inferred and true tumor purity, and scores using Scanpy scoring in the original single-cell data.

Suppl. Figure S9. Average Pearson's correlation across 20 runs and a range of *k* between CDState-inferred and true tumor purity of 33 cancer types from TCGA.

Suppl. Figure S10. Median number of malignant cell states identified across 20 runs and a range of *k* of CDState deconvolution of TCGA datasets. Error bars represent the smallest and the highest number of identified malignant cell states for given *k*.

Suppl. Figure S11. 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.

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 hallmarks of cancer expressed in malignant cell states across TCGA cancer types.

Supplementary Table S10. Recurrent hallmark groups expressed in malignant cell states across TCGA cancer types.

Additional Data

Additional File 1.csv
CDState-inferred proportions of malignant cell states across TCGA cancer datasets.

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. 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.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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: 2285.

24. 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.

25. 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.

26. 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.

27. 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.

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

29. 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.

30. 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.

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

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

33. 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.

34. 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.

35. Kinnersley B, Sud A, Everall A, Cornish AJ, Chubb D, Culliford R, et al. Analysis of 10,478 cancer genomes identifies candidate driver genes and opportunities for precision oncology. Nat Genet. 2024;56: 1868–1877.

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

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

38. 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

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