High-resolution digital dissociation of brain tumors with deep multimodal autoencoder

Jiao Sun[1], Yue Pan[1], Tong Lin[1], Kyle Smith[2], Arzu Onar-Thomas[1], Giles W. Robinson[3], Wei Zhang[5], Paul A. Northcott[2,4], and Qian Li[1*]

[1]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, 38105.
[2]Department of Developmental Neurobiology, St. Jude Children's Research Hospital, Memphis, TN, 38105.
[3]Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN, 38105.
[4]Center of Excellence for Neuro-Oncology Sciences, St. Jude Children's Research Hospital, Memphis, TN, 38105.
[5]Department of Computer Science, University of Central Florida, Orlando, FL, 32816.

*Corresponding author(s). E-mail(s): Qian.Li@STJUDE.ORG;

Abstract

Single-cell technologies enable high-resolution, multi-dimensional analysis of molecular profiles in cancer biology but face challenges related to low coverage and cell annotation. The inherent heterogeneity and complexity of brain tumors may hinder large-scale single cell multi-omic profiling. An efficient alternative is digital dissociation, which involves quantifying cell abundance and purifying bulk samples at high resolution. However, most existing tools for resolving bulk transcriptomes using scRNA-seq as a reference are not easily transferred to other omics (e.g., chromatin accessibility, DNA methylation, protein) due to ambiguous cell markers. Here, we introduce MODE, a novel multimodal autoencoder neural network designed to jointly recover personalized multi-omic profiles and estimate multimodal cellular compositions. MODE is the first algorithm trained on large-scale pseudo-bulk multi-omics derived from an external scRNA-seq reference and an individualized non-RNA reference panel constructed from target tumors. The accuracy of MODE was evaluated through an extensive simulation study, which generated realistic multi-omic data from distinct tissue types. MODE outperformed existing deconvolution pipelines with superior generalizability. Additionally, the high-resolution multi-omic profiles purified by MODE showed strong fidelity and enhanced power to detect differentially expressed genes. We applied MODE to bulk methylome-transcriptome data from two independent brain tumor cohorts, revealing modality-specific cellular landscapes in pediatric medulloblastoma (MB) and adult glioblastoma (GBM). In MB tumors, MODE accurately predicted the composition of embryonal lineage cells and their marker genes expression. In GBM, the deconvoluted cellular composition revealed an increased myeloid cell abundance associated with poorer event-free survival. Overall, MODE high-resolution multimodal dissociation unravels the origins, evolution, and prognosis of pediatric and adult brain tumors, offering a powerful tool for multi-omic analysis at cell state resolution without large-scale single cell sequencing. MODE is a pipeline freely available at https://github.com/jsuncompubio/MODE.

Keywords: multimodal, autoencoder, high-resolution purification, tumor microenvironment

Main

Multimodal omics profiling and analysis, including transcriptomics, epigenomics, and proteomics, are crucial in cancer and immunology research. Initiatives like The Cancer Genome Atlas (TCGA) project and Clinical Proteomic Tumor Analysis Consortium (CPTAC) have performed bulk multi-omics analysis on various cancer types [1, 2]. Recently, single-cell RNA sequencing (scRNA-seq) combined with other (epi)genomic data has enabled a deeper understanding of intra-tumoral heterogeneity, the identification of rare cell populations, and the exploration of tumor microenvironment interactions [3–6]. The integration of transcriptomic, epigenomic, and proteomic data provides insights into immune cell states and their functional responses to stimuli, such as therapeutic interventions, facilitating a better understanding of tumor evolution and the identification of potential therapeutic targets.

Emerging technologies in single cell (multi-)omics like scATAC-seq, cyTOF, CITE-seq (protein and gene expressions) and single cell multiome (ATAC + gene expression) [7–10] offer promising avenues for the multi-dimensional integration at single cell resolution, enabling holistic understanding of complex biological processes. A well-known challenge in single cell multi-omics data preprocessing is the cell marker selection and cross-modality cell annotation. Tools like Seurat [11–14] and MaxFuse [15] have introduced innovative methods for cross-modality cell annotation in scATAC-seq or snmC-seq3 [16] based on reference scMultiome. These approaches may not be ideal for tumor tissues due to unexpected intra-tumoral and inter-tumoral heterogeneity. A rare cell state in target tumors may not be present in a reference scMultiome data, which can play a critical role in disease prognosis and treatment response. Additionally, for population-level analyses involving large cohorts, single cell multi-omic profiling may not be scalable due to the sequencing costs and computational complexity.

Another challenge in single-cell analysis arises in the lineage mapping of cells in embryonal tumors, such as pediatric medulloblastoma (MB). Multiple studies have employed machine learning and deep learning algorithms to assign each cell in MB tumors to the closest matching cell state in a reference human embryonal cerebellum tissue [17, 18]. However, in Group 3 MB tumors, many cells were not successfully assigned to any predefined lineage cell state due to significant differences between the reference embryonal normal tissue and the target tumors. Moreover, this type of analysis may not be applicable to epigenomic or regulatory landscapes (i.e., DNA methylation and chromatin accessibility) because of the sparseness in single cell data and the aforementioned problems in preprocessing.

An alternative approach for dissociating bulk multi-omics data in large cohorts, without the need for single-cell sequencing and annotation, is to digitally deconvolute and purify bulk samples using external single-cell references. Most deconvolution methods [19–22] assume that the target bulk data and reference data share similar biological information at the cell state level, i.e., a signature matrix. If the target and reference datasets are generated from distinct platforms, normalization procedures

can reduce discrepancies and enhance the accuracy of the output [23]. A more advanced solution for addressing platform-related disparities is to train deep learning models, such as Scaden [24], TAPE [25], or scpDeconv [26], on pseudo-bulk replicates derived from reference scRNA-seq data, and then adapt the trained models to each target bulk sample. But limitations still exist in these tools. First, to resolve the epigenome or proteome in the absence of a modality-matched single-cell reference, scRNA-seq data with linked cell markers is necessary, such as deconvoluting bulk proteomes using a scRNA-seq reference in scpDeconv [26]. If such references are unavailable, users must provide non-RNA single-cell references with preprocessed cell labels, such as single cell mass spectrometry [27] or cyTOF proteomics [8], scATAC-seq, bisulfite sequencing DNA methylation [28, 29], whereas the cell marker coverage and cell annotation are ambiguous and challenging. Second, the accuracy of differential (expression) feature detection between disease subgroups from the digitally purified profiles has not been rigorously validated with experimental benchmarks for existing tools (e.g., CIBERSORTx, TAPE), potentially leading to inflated false positives. Lastly, current deep learning deconvolution tools are trained using pseudo-bulk data derived from external references, which may not capture sufficient between-sample variability relative to the target bulk data. One concern is that deep learning algorithms may occasionally perform poorly on target data due to unexpected characteristics (or randomness) not captured during the training phase. Thus, deep learning models trained on external and limited scRNA-seq references may fail to accurately predict cellular composition and high-resolution profiles in target tissues. To improve model performance, it is essential to incorporate the variability among target tissues into the training process.

A recent study employed the MuSiC pipeline [21] to trace the origins and evolution of pediatric medulloblastoma (MB) tumors by deconvoluting bulk RNA-seq data using a signature matrix derived from human fetal cerebellum normal tissues [17]. However, this approach did not account for the tumor-normal differences, potentially underestimating the contribution of cellular subpopulations in the tumor's origin. In another study of adult glioblastoma multiforme (GBM), bulk epigenetic profiles were deconvoluted using EpiDish [30], revealing differences in myeloid cell abundance across patient subgroups with distinct survival outcomes [31]. These subgroups were classified based on epigenetic signatures of neural cells, but the deconvolved fractions of myeloid or neuron cells were not directly linked to survival outcomes. This suggests that deconvolution of DNAm using external reference panels may not fully capture the latent inter-tumoral heterogeneity.

To address the challenge of dissociating bulk epigenomes paired with transcriptomes, particularly when features (e.g., gene-peak or gene-CpG) are weakly linked, we propose a MultimOdal Deep autoEncoder (MODE) framework. MODE leverages cell markers detected in scRNA-seq data and accounts for the reference-to-target disparity, without requiring cell annotation across molecular modalities. To illustrate and benchmark the accuracy of MODE, we designed an extensive simulation study using the multimodal single-cell simulator scDesign3 [32–34], and compared its performance to other

deep learning tools such as Scaden [24], TAPE [25], and scpDeconv [26]. In both tumor and normal tissues, MODE consistently outperformed the competing methods, accurately quantifying cellular fractions in RNA and non-RNA molecules. We further assessed the sensitivity and specificity of differential expression (ctsDE) analysis by applying the Wilcoxon Rank-Sum test to the digitally purified profiles in normal tissues. Our results demonstrated that MODE predictions improved the accuracy of downstream ctsDE detection by leveraging multimodal omics data. Finally, when applied to multiple brain tumor studies, MODE successfully uncovered distinct lineage cell states involved in the evolution of pediatric MB tumor subgroups and benchmarked the prognostic value of the GBM tumor microenvironment composition across molecular modalities.

Results

Overview of MODE

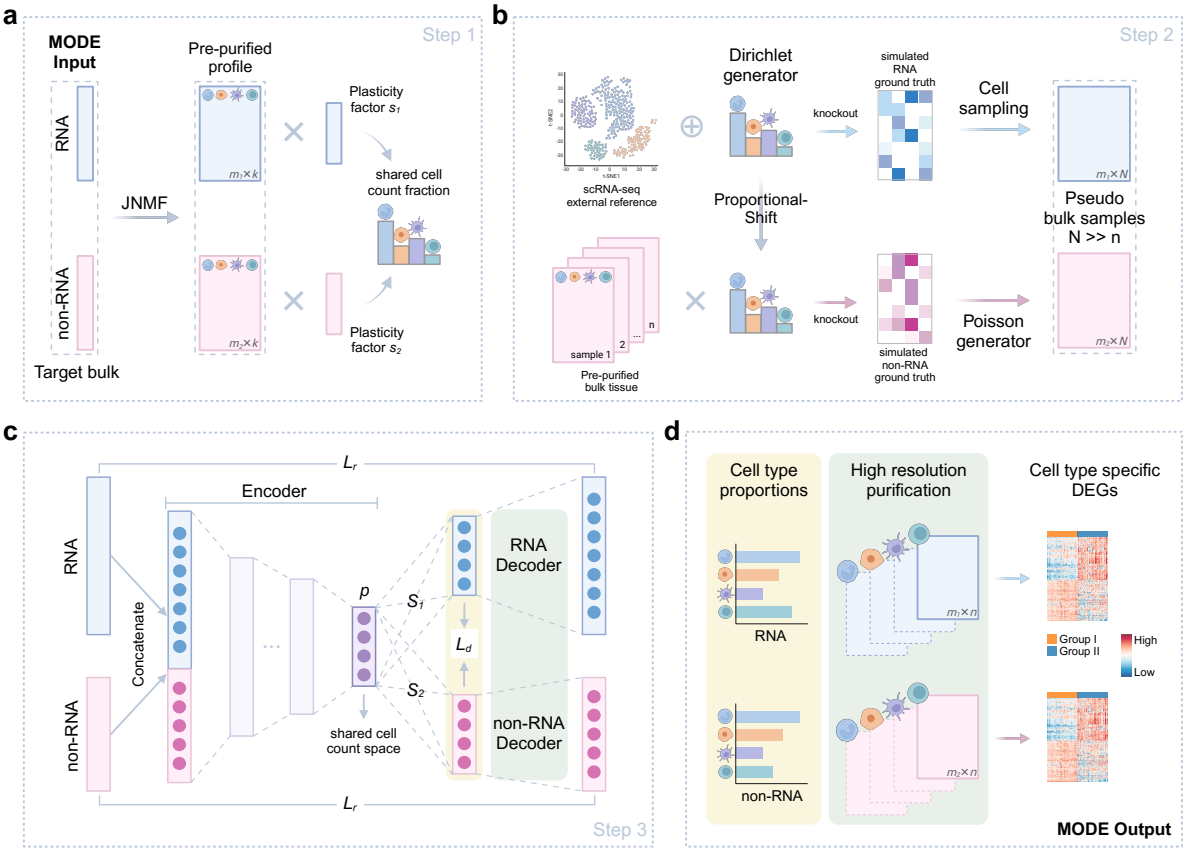

Fig. 1 Overview of MODE framework. (a) The input consists of RNA and non-RNA bulk omics data paired by tissues. MODE initializes a personalized signature matrix via Joint Nonnegative Matrix Factorization and shared cell counts fractions. **(b)** Generate large-scale training data based on the input scRNA-seq reference and the personalized signature matrices from Step 1. A proportional shift strategy coupled with modality-specific sparsity is used to simulate ground truth cellular compositions in RNA and non-RNA molecules. **(c)** A multimodal autoencoder with one encoder for the concatenated input data and two decoders for reconstructed bulk multiome. An updated personalized multi-omics signature matrix is extracted from two decoders adapted to each target tissue. **(d)** MODE outputs the predicted cell proportions and high-resolution purified multi-omic profiles, enabling downstream cts-differential analysis.

We present MODE, a semi-reference based multi-omic digital dissociation method that consists of three steps: (1) pre-purification for target bulk multi-omics data; (2) generation of semi-reference free multi-modal pseudo-bulk training data; (3) multimodal autoencoder adapted to target tissues, as illustrated in Figure 1. In Step 1, we initialize the tissue-specific reference panels for the target bulk non-RNA profiles (i.e.,epigenome or proteome) via the Joint Non-negative Matrix Factorization (JNMF) model and Projected Gradient Descent optimization. This model links the paired bulk transcriptome and non-RNA profiles through tissue-specific cell counts fractions and generated pre-purified multiome as output, as illustrated in our previous work [35]. In this step, we use scaling factors $(\boldsymbol{s}_1, \boldsymbol{s}_2)$ to account for the modality-specific cell plasticity (Figure 1a).

The second step of MODE (Figure 1b) aims to generate large-scale pseudo-bulk multiome data by connecting different molecular modalities through a proportion-shift strategy. Specifically, we first simulate the ground truth subcellular proportions in transcriptome by Dirichlet distribution and *a priori* parameters. An external scRNA-seq reference with preprocessed cell labels is used to sample the cells per ground truth proportions, which are aggregated as a pseudo-bulk transcriptome. The usage of scRNA-seq reference ensures the gene expression heterogeneity across cellular subpopulations and pseudo samples. For the epigenome or proteome paired to transcriptome, the ground truth proportions are generated by shifting RNA proportions, mimicking the shared cellular landscape and introducing cross-modality variability. The non-RNA cellular fractions and the tissue-specific purified data recovered in Step 1 are then used to simulate the non-RNA pseudo-bulk profiles, with between-sample variability further enhanced by Poisson random generator. This process is guided by the target bulk non-RNA profiles, as well as leverages the inter-cellular activity from an external scRNA-seq data. The randomness in training data introduced by Step 1 target tissue purification and Step 2 probabilistic data generation process ensures our model being optimized by a large pool of pseudo samples spanning the latent heterogeneity in the external reference and target tissues.

In Step 3 (Figure 1c), we develop a multimodal autoencoder for tissue-adaptive deconvolution which integrates RNA and non-RNA bulk data to learn the shared cellular space. It consists of one encoder for the concatenated bulk multi-omics input and two parallel decoders for reconstructing the input bulk data per modality. The encoder reduces the concatenated high-dimensional bulk expression profiles to a low-dimensional embedding $(\boldsymbol{p})$ for cellular landscape shared between molecular sources. This latent space representation is then linearly transformed (via $\boldsymbol{S}_1, \boldsymbol{S}_2$) into the modality-specific cellular composition and the reconstructed bulk multiome. This framework is trained with the simulated bulk multi-omics data in Step 2 to minimize the loss in deconvoluted cellular composition and reconstructed bulk data for both modalities. In the adaptive stage, the parameters are refined for each target bulk sample to better capture the unseen patterns present in the tissue. It is worth to note that the feature selection in bulk epigenome or proteome does not require cell markers identified from non-RNA single cell reference.

Instead, users can select the features from bulk data by routine metrics, such as feature-wise mean, variation, and zero-inflation.

Tumor tissue gene expression and chromatin accessibility simulation

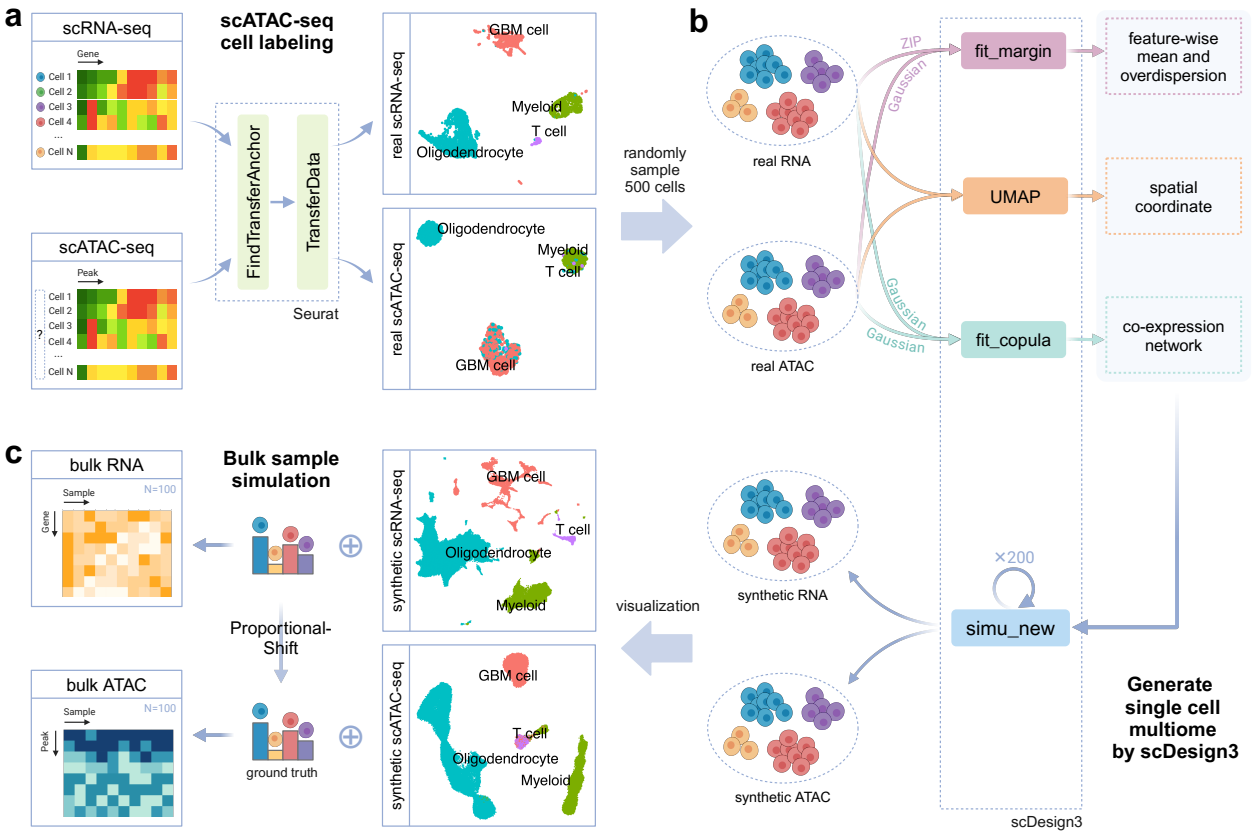

Fig. 2 Tumor tissue paired multiome simulation. (a) Transfer scRNA-seq cell labels to scATAC-seq by Seurat. **(b)** Synthetic single cell multiome generation by scDesign3. **(c)** Simulated pseudo bulk gene expression and chromatin accessibility.

We designed a novel simulation study with scDesign3 [34] to compare different methods for the deconvolution of bulk multiome in solid tumors. To mimic the variability among tumor tissues, we utilized the real scRNA-seq and scATAC-seq data paired by GBM tumor samples [36] to estimate the statistical parameters required by scDesign3 simulator, visualized in Figure 2. The cell labels of this real scATAC-seq data were predicted by Seurat [13], which transferred the scRNA-seq cell labels to scATAC-seq in the same biological system. Although this multimodal cell annotation algorithm cannot label some GBM malignant cells precisely in scATAC-seq according to paired UMAPs (Figure. 2a), it does not affect the overall statistical parameters extracted from gene expression and chromatin accessibility per cell type in the next step (Figure. 2b). The synthetic single cells per modality generated from scDesign3 are repeatedly sampled and aggregated as psuedo bulk RNA-seq and ATAC-seq profiles, as shown in Figure. 2c. The details of this simulation process are described in Methods.

We compared MODE with the competing methods (scpDeconv, TAPE, and Scaden) by generating target data with small vs. large difference across modalities in cell fractions (Figure 3a). The cell marker genes were identified from real scRNA-seq data in a previous study [36], while the features (peaks) in scATAC-seq detected in $> 10\%$ cells were selected. TAPE proposed tissue-adaptive cell deconvolution on bulk RNA-seq with scRNA-seq as reference, while scpDeconv is designed for proteomics deconvolution with missing feature imputation and domain-adversarial training strategies. For each target data, MODE can accurately estimate the RNA-seq and ATAC-seq cell proportions. All the competing methods can deconvolve bulk transcriptome with scRNA-seq reference, in which TAPE and Scaden achieved performance comparable to MODE with Lin's Concordance Correlation Coefficient (CCC) > 0.9. However, TAPE, Scaden, and scpDeconv failed to deconvolve pseudo bulk ATAC-seq even with the ground truth scATAC-seq data as reference due to the sparseness in counts and lack of cell marker features in chromatin accessibility. The tool scpDeconv developed for proteomics deconvolution was the second-best performer in ATAC-seq cell abundance quantification with CCC $= 0.432, 0.587$. Additionally, for the competing methods, we assessed the concordance between the deconvoluted RNA proportions vs. ground truth ATAC-seq proportions as surrogate estimate. In the scenario of small difference between RNA and ATAC, the result (Supplementary Figure S1) implied that RNA cellular fractions resolved by TAPE (CCC $= 0.88$) and Scaden (CCC $= 0.93$) were not superior to MODE ATAC-seq fractions (CCC $= 0.92$). It's worth noting that Scaden is a tool only focusing on cell abundance quantification without digital purification, although its performance in pseudo bulk RNA-seq cell abundance quantification is equivalent to MODE. Hence, RNA proportions cannot serve as surrogate for non-RNA proportions if the cellular composition varies tremendously across molecular sources.

Normal tissue transcriptomics and proteomics simulation

To evaluate the sustainability of MODE, we designed another experiment to generate nine scenarios of bulk RNA-seq transcriptome and mass spectrometry proteome for human brain prefrontal cortex normal tissues, using public snRNA-seq data [37] and pure cell proteomes [38]. Each scenario consisted of a target bulk multi-omic dataset defined by a sample size $N = 50, 100, 200$ and cts-differential expression (ctsDE) log fold change (LFC$= 0.5, 1, 1.5$) between diseased and controls. Details of this simulation design were described in previous work [35]. We compared MODE with scpDeconv, TAPE, and Scaden (Figure 3b and Supplementary Figure S2). For each target bulk multi-omic data, we reran MODE by permuting different hyperparameters, such as initial cell fraction variance, step size(s) and convergence threshold in Step 1, sparseness probability in Step 2, and batch size in Step 3. TAPE and Scaden had poor performance with average CCC$= 0.020$ and -0.112 across all scenarios, revealing that these methods cannot handle the significant difference between signature matrix and target bulk profiles. The domain-adversarial training in scpDeconv narrowed the gap between reference scRNA-seq and target

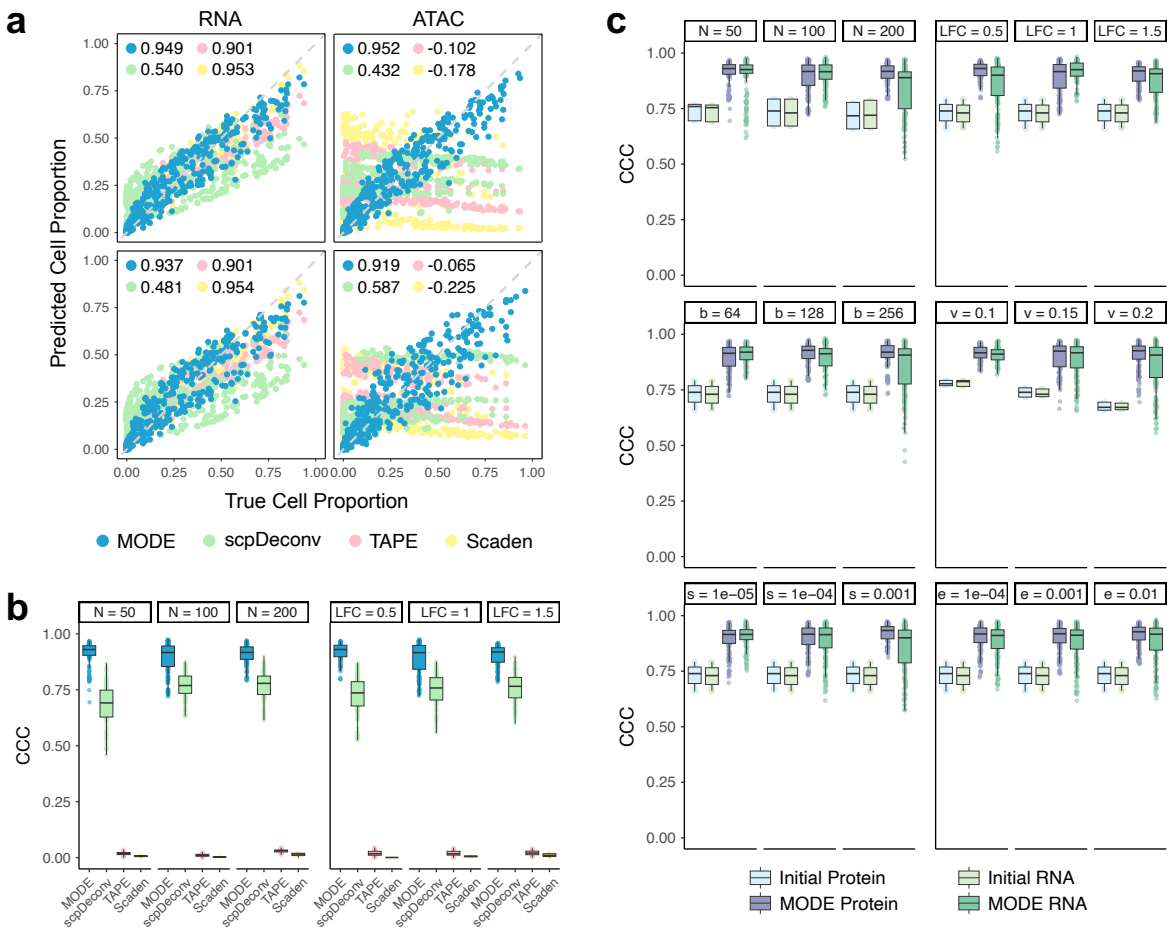

Fig. 3 Performance comparison of MODE and competing methods on simulated tumor and normal tissue data. (a) Scatter plots of true (x axis) and predicted cell-type proportions (y axis) by deconvolving tumor simulated bulk gene expression and chromatin accessibility with either scRNA-seq or scATAC-seq from GBM tumor as reference. There are two scenarios representing small (top) and large (bottom) difference between molecular sources. The CCC between predicted proportion and ground truth is shown on the top left of each plot. **(b)** CCC of MODE, scpDeconv, TAPE, and Scaden on simulated normal proteome deconvolution. **(c)** CCC of MODE and the initial cell count fractions in Step 1 JNMF purification by different normal tissue simulation scenarios: sample size (N), log fold change (LFC); MODE hyperparameters: batch size (b), initial value variance (v), step size (s), and convergence threshold (e).

bulk protein data with reasonable deconvolution performance (i.e., mean CCC= 0.738), while MODE in contrast consistently outperformed TAPE, scpDeconv, and Scaden on bulk protein deconvolution. MODE has demonstrated superior performance of cell abundance quantification in both omics across all scenarios and hyperparameters, substantially improving the accuracy compared to the initialized cell counts fractions in terms of CCC and MAE (Figure 3c and Supplementary Figure S3). Specifically, MODE yielded stable cell fraction estimates regardless of the variation in initial values. Our multimodal autoencoder remained powerful for a wide range of hyperparameters in the JNMF initialization and parallel decoder training, serving as a tuning-free deconvolution pipeline.

In the ground truth of astrocyte- and neuron-specific multi-omics expression data, there are 100 features per modality generated as ctsDE for each cell type. We evaluated the quality of high-resolution ctsMultiome profiles purified by MODE and TAPE based on the accuracy of downstream ctsDE detection. For the purified data in each molecular source, features with zero variance across samples were

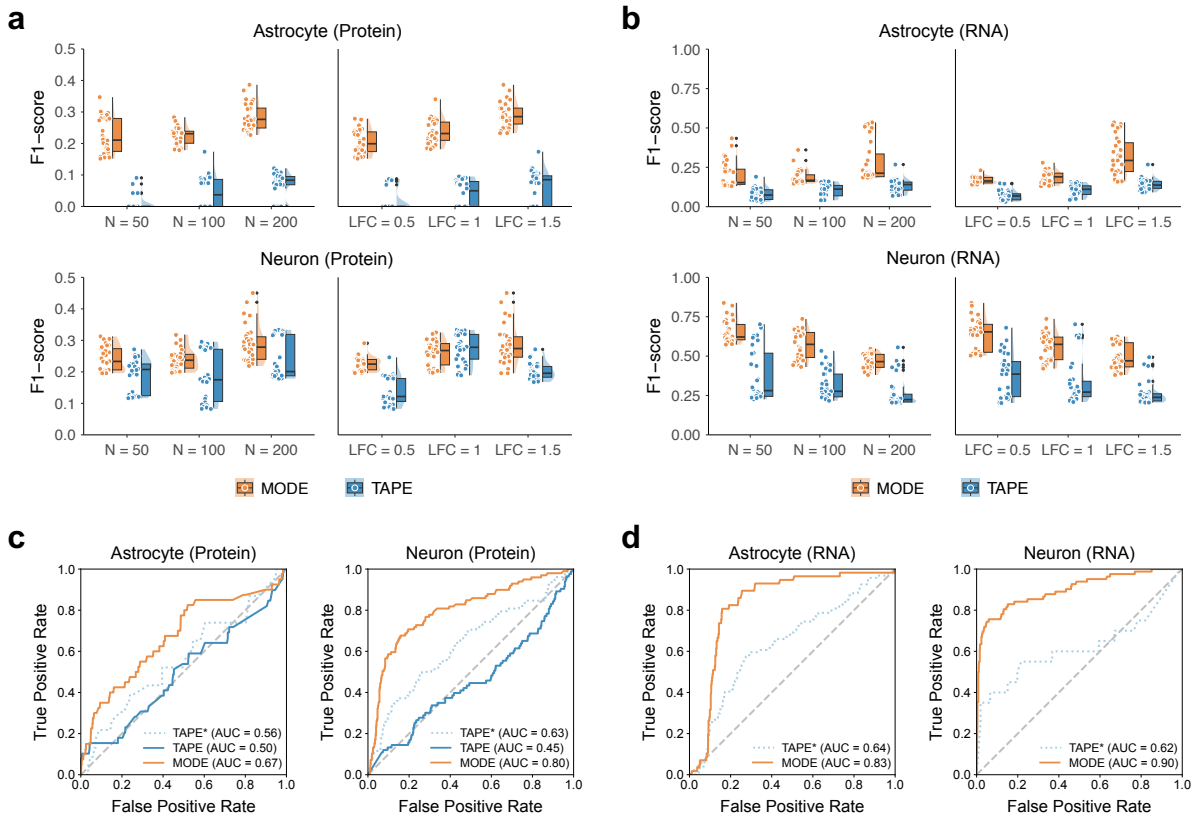

Fig. 4 Cell-type-specific multiome profiles recover differentially expressed features. (a, b) Raincloud plots for F1-score on ctsDE detection in Astrocyte- and Neuron-specific proteome and transcriptome by MODE and TAPE. **(c, d)** Area Under the Receiver Operating Characteristic Curve (AUC) for the best F1-score replicate by each method based on the same target data (N=200, LFC=1.5). Solid line represents the hyperparameter setting used in the best result of MODE, which may differ from or be identical to the default setting used in TAPE (dash line).

considered invalid and excluded in the downstream DE analysis. We used Wilcoxon Rank-Sum test to identify DE features in cts-transcriptome and -proteome, individually, with false discovery rate (FDR) corrected by Benjamini-Hochberg (BH) procedure. Features with FDR-adjusted $p < 0.05$ were called as DE. We used sensitivity, specificity, and the F1-score between ground truth vs. predicted labels as metrics to evaluate the power and robustness of DE detection.

Among the replicates of high-resolution multiome purified by distinct hyperparameters, we filtered out those with less than 300 valid features and summarized the overall performance on the remaining replicates. The best F1-score per method in purified transcriptomics and proteomics were summarized for each simulation scenario in Figure 4a-b, demonstrating superior performance of MODE on ctsDE detection across different sample sizes and log fold change levels. It's worth noting that as sample size or LFC increases, the Wilcoxon test on MODE-purified data yields better F1-score for both protein and RNA expressions, while the accuracy of DEG detection based on TAPE-purified RNA-seq profiles does not increase at larger sample size and LFC. The results in proteomics purification demonstrated that MODE was robust and powerful in predicting protein expression regardless of cell fractions, while the quality of TAPE-purified proteomics data was unstable based on the random change of accuracy in

downstream DE analysis (Figure 4a). On the other hand, the neuron-specific RNA-seq gene expression recovered by either method yields inflated false positive rate (Supplementary File 2) with lower F1 score at larger sample size and LFC (Figure 4b). This result in neuron-specific DEG detection implied that digital purification for the most abundant cell state (e.g., neuron cells) through deep autoencoder methods may overestimate the variability in RNA-seq data. Meanwhile, we selected the results of MODE and TAPE from similar or identical hyperparameters for the scenario of maximum sample size (N=200) and LFC (1.5). The Area Under the Curve (AUC) illustrated significant improvement by MODE on DE detection in astrocyte and neuron cells across modalities, visualized in Figure 4c-d. In general, MODE substantially improved the quality of high-resolution ctsMultiome by leveraging non-RNA profiles in the target tissues.

Application to pediatric medulloblastoma

Medulloblastoma is one of the most malignant childhood brain tumors with diversified treatment responses and stratified outcomes. A recent study [17] mapped pediatric MB tumors to the human fetal cerebellum origins by deconvolving the bulk transcriptome of tumor tissues collected in the International Cancer Genome Consortium (ICGC) [39] and the SJMB12 cohort. SJMB12 is a multi-center clinical trial enrolling 660 children ages 4-25 with newly diagnosed medulloblastoma. Importantly, this trial was the first to incorporate molecular subgrouping into risk stratification, tailoring therapy to both biological and clinical features.

In this analysis, the deconvolution tool MuSiC [21] and a scRNA-seq reference data from human fetal cerebellum normal tissues [40] were deployed to deconvolve bulk RNA-seq data from these cohorts [17]. The estimated cell composition of the transcriptome was found to correlate with MB subgroups clinically classified based on DNAm signatures: Sonic Hedgehog (SHH), Group 3 (G3), and Group 4 (G4). However, cell types such as Astro/Glia and GlutaCN/UBC—identified as progenitors or the lineage of origin for G3 and G4 MB malignant cells in scRNA-seq analyses [41]—were not fully recovered by MuSiC in the bulk transcriptome due to computational limitations. Meanwhile, paired methylome-transcriptome data from both cohorts have been previously profiled with detailed preprocessing descriptions [17, 42].

To illustrate the performance of our new method, we applied MODE to the paired bulk multi-omics of N=253 MB tumors, using the above human developmental cerebellum scRNA-seq data as reference. We used 3146 cell type marker genes identified from this reference and selected 1000 DNAm features contributing to the cell type heterogeneity and latent cellular embeddings shared across modalities. These features were chosen from the most variable CpGs linked to cell marker genes, using MethylMix [43] and MICSQTL [35]. The deconvoluted cell fractions and purified gene expression were compared between MB subgroups using the Wilcoxon Rank-Sum test and BH correction for FDR.

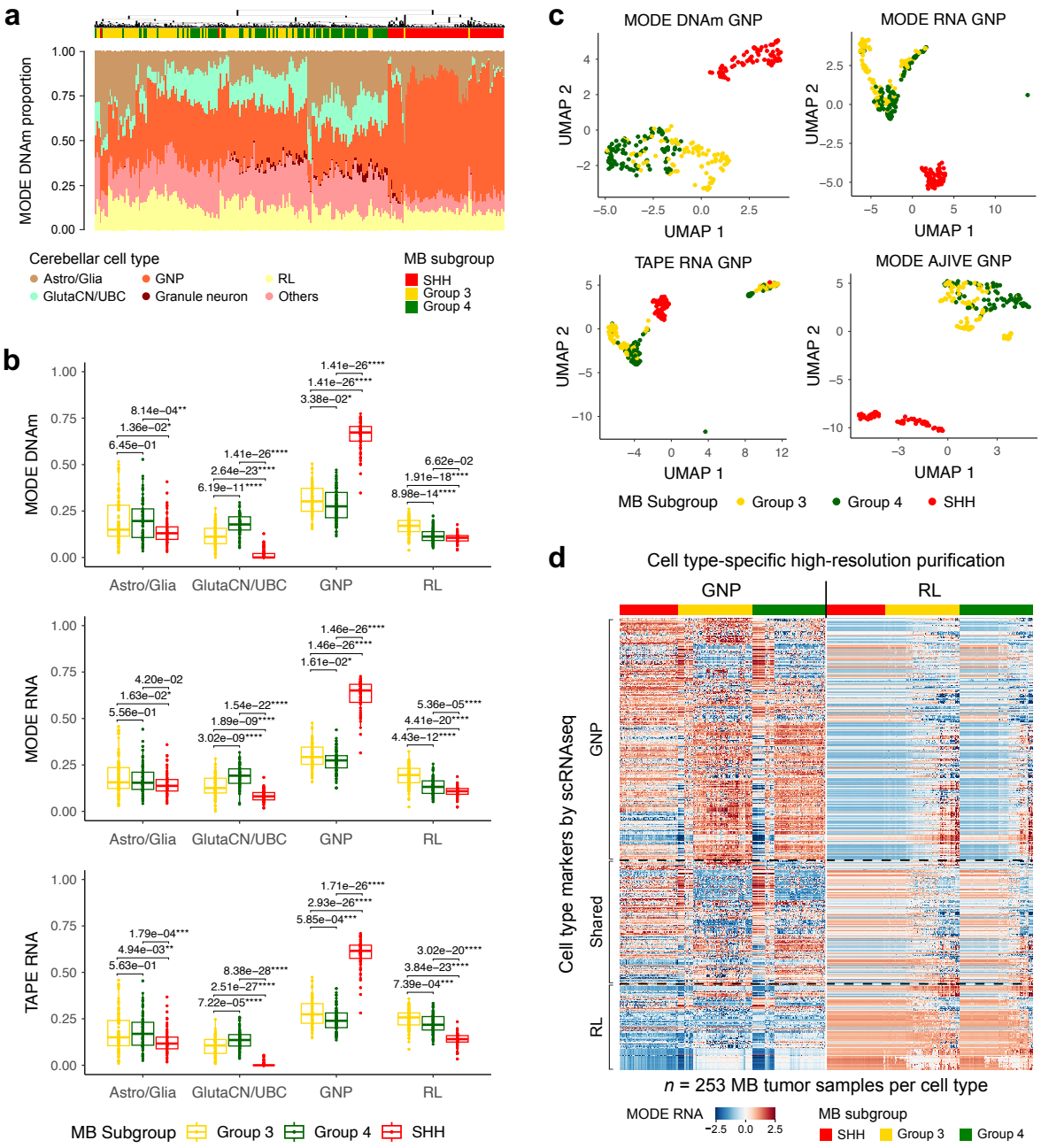

Fig. 5 Multi-omic dissociation of pediatric medulloblastoma tumors with human fetal cerebellum scRNA-seq reference. (a) Origin cell proportions in bulk MB methylomes quantified by MODE. **(b)** Comparison of MODE (DNAm, RNA) and TAPE (RNA) cell type proportions across MB subgroups. **(c)** UMAP of MODE (DNAm, RNA), MODE multiomes by AJIVE, and TAPE (RNA) purified profiles mapped to GNP cell state. **(d)** MODE-predicted MB tumor gene expression mapped to GNP and RL lineages for their marker genes identified in scRNA-seq data of human embryonal cerebellum.

The human fetal cerebellum derived cell states mapping to MB tumors were quantified by MODE in both methylome (Figure 5a) and transcriptome (Supplementary Figure S4), outperforming TAPE (Supplementary Figure S5). Both MODE and TAPE predicted a greater fraction of granule neuron progenitor cells (GNPs) and rhombic lip (RL) in SHH-MB and G3/4-MB, respectively, based on either gene expression or DNAm (Figure 5b). Additionally, for each molecular modality, G3-MB patients had the most abundant RL fraction compared to the other two subgroups ($p < 10^{-11}$), while G4-MB tumors

were more likely enriched in Unipolar Brush Cells (GlutaCN/UBC) ($p < 10^{-8}$). Notably, the cellular fractions resolved by MODE revealed that GlutaCN/UBC origin cells were more abundant than RL in G4-MB tumors, while the opposite was observed in G3-MB, as a distinction not captured by TAPE. The differences between G3 vs G4 in RL stem cells and more mature GlutaCN/UBC proportions in G4 were more pronounced in MODE predictions, supporting the lineage trajectories for G3/4 MB tumors identified in a previous single-cell analysis [17]. The higher abundance of GNPs in the SHH subgroup, quantified by both MODE and TAPE, further demonstrate the role of GNPs in SHH pathway activation during MB tumorigenesis in GNPs [44].

The digitally purified MB tumor methylation and gene expression also demonstrated superior power of MODE. DNAm beta values for select CpGs and expression levels for the lineage-specific genes in MB tumors were successfully assigned to the GNP subpopulation and visualized by UMAP (Figure 5c), including a joint visualization of MODE purified multi-omic profiles using the AJIVE algorithm [35, 45]. The quality of purified profiles in the other cell states (using either MODE or TAPE) was not ideal for UMAP visualization due to the low coverage of predicted features. Notably, GNP-only markers are highly expressed within the GNP subspace projected to MB tumors, while RL-only markers show overall higher expression in the RL subspace (Figure 5d).

We further assessed the DEGs for MB subgroups based on purified data for GNP and/or RL markers, identifying 446 GNP marker genes and 175 RL marker genes differentially expressed ($p < 0.05$) among MB tumor subgroups derived from each developmental cell state (Figure 5d). Several genes previously reported as subgroup-specific drivers or tumor lineage markers, such as *PTCH1, PDE1C, ST18* in SHH [18, 42] and *OTX2, EOMS, LMX1A* in G3 [18, 46], were recovered and detected as DEGs exclusively in GNP and RL, respectively. This result aligns well with the fact that GNP and RL are the embryonal origins for SHH and G3 MB tumors. Moreover, the GNP-specific DNAm profiles predicted by MODE (Supplementary Figure S7) showed clear discrimination between SHH and G3/4 with strong DE patterns, as well as accurately classified most of G3 vs. G4 patients through hierarchical clustering.

Application to TCGA adult glioblastoma

To further investigate the deconvolution power and robustness in median-size cohorts, we applied MODE and TAPE to TCGA Glioblastoma multiforme (GBM) data. This dataset contains bulk RNA-seq and DNA methylation paired by the tumors from $N = 64$ GBM patients. The RNA-seq transcript expression in transcript per million (TPM) was downloaded from UCSC Xena [47], and then aggregated to gene-level expression by GenCode annotation. The DNA methylation beta values was downloaded from the GDAC data portal https://gdac.broadinstitute.org/. We used a preprocessed and annotated scRNA-seq data from GBM patients in an external cohort [5, 36] as reference to resolve the microenvironment of bulk transcriptome-methylome with six cell types (astrocyte, GBM cancer cell, myeloid, neuron,

oligodendrocyte, and T cell). There are 2115 DEGs identified as cell markers from this scRNA-seq data [36]. We selected the top 10000 variation CpGs and then mapped the probes to DEGs using MethylMix [43], resulting in 965 CpGs. The clinical data of patients were collected from cBioportal [48]. The Kaplan Meier (KM) curve and log-rank test analysis were conducted based on the high- and low- cell abundance groups stratified by the binarized estimated cell proportions.

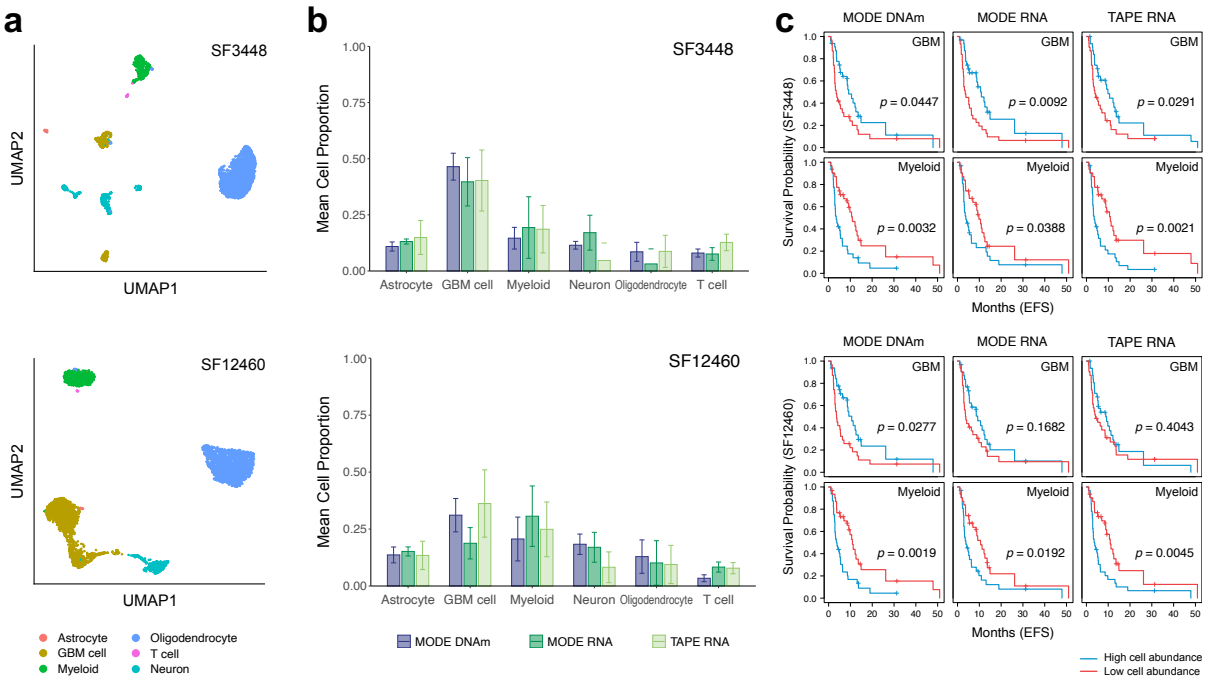

Fig. 6 Dissociation of TCGA GBM bulk multi-omics data with different scRNA-seq references. (a) Uniform Manifold Approximation and Projection (UMAP) visualizations of an external scRNA-seq reference data collected from two patients. **(b)** The mean cellular compositions estimated by MODE (DNAm, RNA) and TAPE (RNA) with single cell reference in **(a)**. **(c)** TCGA GBM patients' survival outcomes stratified by binarized cell proportions.

The scRNA-seq reference data was profiled from two GBM patients with identifiers SF3448 and SF12460, individually, as visualized in Figure 6a. Both MODE and TAPE use the default sparsity parameter at 0.5 to evaluate the impact of reference data on each method. Obviously, the confidence intervals per cell type in (Figure. 6b) exhibits significant difference between methods and molecules in the context of prediction uncertainty. Specifically, MODE yields the smallest variation of cellular fractions, especially for astrocyte abundance in either types of molecules. The Neuron cell fraction was stably recovered by MODE from the bulk multi-omics data, while TAPE failed to quantify this subpopulation in most target tissues with reference SF34348, in which the neuron cells were not well clustered. On the other hand, the estimated abundance of oligodendrocyte in RNA was not ideal in either method, but MODE was able to recover this cell state in the bulk DNAm profiles. The overall distribution of the resolved tumor microenvironment based on either reference aligns with our hypothesis that MODE was not severely affected by the cell annotation quality in scRNA-seq reference because of the target tissue-recovered reference panel used in the training process.

We further examined the predicted microenvironment composition by linking it to the survival outcomes of patients who were classified as high- and low- cellular subgroups based on the median abundance per cell type. Both MODE and TAPE uncovered that higher myeloid abundance was associated with poor event free survival (EFS) across different deconvolution signature matrices (see Supplementary Table. S1), being consistent with a recent finding about myeloid in GBM [31]. For the reference SF3448, MODE RNA proportion of GBM tumor cells ($p = 0.0092$) presented prognostic detection power comparable or slightly higher than TAPE RNA result ($p = 0.0291$) (Figure 6c). Meanwhile, the GBM proportion in methylome estimated by MODE with reference SF12460 revealed clear EFS stratification among patients ($p = 0.0277$), while TAPE cannot identify this difference in RNA with the same reference($p = 0.4043$). This result also implied that higher tumor purity in GBM tissues may delay the recurrence or progression of this disease. Similarly, by using the same reference of SF12460, the neuron and T cell fractions in DNAm showed better stratification of EFS (Supplementary Table. S1), individually, which was not observed in TAPE RNA fractions or previous GBM deconvolution analyses. The log-rank test results for other cell states and overall survival outcome were shown in Supplementary Figure S8 and Table S1. In summary, by recovering the pure cell bulk methylome from target tumors, MODE unravelled the tumor microenvironment accurately with higher prognostic detection power than TAPE.

Discussion

We developed a multimodal deconvolution method MODE that captures cellular processes between molecular modalities. Through JNMF initialization, MODE recovered the purified profiles from bulk epigenome or proteome and provided a personalized reference panel to build pseudo bulk samples for training. The multimodal autoencoder network in MODE jointly quantifies the cell abundances in RNA and non-RNA molecules through a shared cell count representation and yields cell-type specific profiles per modality. The MODE framework achieved multi-omics joint learning by using proportional shift to generate sample-matched pseudo bulk data for training. Systematic and comprehensive experiments demonstrated MODE's superior performance and robustness on multi-omics cell deconvolution with an emphasis on non-RNA omics. MODE's deconvolution performance is stable and robust across scenarios and experiment setups as demonstrated on normal simulation datasets. On both simulated normal and tumor tissue multiome, MODE consistently outperforms state-of-the-art methods in bulk non-RNA deconvolution using only scRNA-seq references. In addition, application to real pediatric MB and adult GBM patient samples highlights the clinical value of MODE in cancer diagnosis and prognosis.

MODE is distinguished with four key advantages compared to the existing single omics deconvolution methods. (1) MODE can accurately estimate non-RNA cell proportions by multimodal joint learning with scRNA-seq reference only, without the requirement of non-RNA cell markers or linked features. (2) MODE's multimodal autoencoder network is trained on the simulated pseudo bulk multiome samples,

of which the non-RNA pseudo bulk is built on top of the purified data recovered from target samples. This approach not only bridges the gap between single cell reference and target bulk data, but also incorporates realistic between-sample heterogeneity into the training data. (3) MODE introduces the proportional shift strategy to effectively link RNA and non-RNA cellular composition in the same sample to enable integrative deconvolution. (4) MODE can recover rare cell states to decipher the complex tumor microenvironment and potentially identify cts-signatures predictive of disease onset and prognosis.

The current work of MODE improved both cell abundance quantification and high-resolution digital purification of bulk multiome compared to TAPE, being a framework generalizable to various omics data. The quality of digital purification by MODE may depend on hyperparameter selection, which can be determined and tuned by a pilot scRNA-seq experiment or *a priori* ctsDE marker genes. In addition, the one-step ctsDE analysis methods like TOAST [49] that identify ctsDEGs based on mixture bulk samples and pre-estimated cellular fractions without purification, can guide hyperparameter tuning for MODE and validate the quality of cell-type-specific purification. This also eliminates the need for pure cell reference profiles, which are often challenging to obtain in brain tumors. One possible extension to MODE is to enhance the consistency in digital purification by leveraging and imputing an external scMultiome reference. Our current algorithm (Step 2) constructs the pseudo bulk multi-omics data from an external scRNA-seq data and the tissue-recovered non-RNA signature matrices, without imputing the missing features in scRNA-seq. Hence, recovering the undetected features in single cell reference may increase the power of predicting individualized ctsMultiome in future. If a multimodal single cell reference is lacking, one may adopt scDesign3 to generate synthetic data that mimics transcriptome-epigenome or proteome paired at cell resolution. Another potential innovation built on MODE is to resolve the spatial and cellular landscapes in tissues by integrating bulk multi-omic profiles, whole slide imaging, and a spatial transcriptomics reference. A recent research reconstructed the spatial landscape of malignant cells in breast cancer bulk transcriptome via graph attention neural network [50]. We may incorporate this approach in the MODE framework to predict multimodal cellular landscape per spot, as well as identify the regional biomarkers not detectable in current spatial technologies. Together, we will be able to elucidate the underlying mechanisms for disease with cross-modality segmentation of biospecimens.

Methods

The MODE framework

For each tumor tissue, we measure the mixed-cell bulk expression of gene g ($g = 1, ..., m_1$) and the abundance of feature h ($h = 1, ..., m_2$) in the other omics (e.g., CpG sites, peaks for chromatin accessibility, proteins), respectively, denoted by y_{1g}, y_{2h}. The latent individualized measurements for cell type c ($c = 1, ..., k$) are denoted by x_{1gc}, x_{2hc}. The modality-specific cellular fractions are θ_{1c}, θ_{2c}, determined

by the common tissue-specific cell counts (fractions) p_c and source-specific cell size factors s_{1c}, s_{2c}. That is $\theta_{1c} = p_c s_{1c}$, $\theta_{2c} = p_c s_{2c}$. Thus, the paired bulk multi-omic data are described as

$$E(y_{1g}) = \sum_{c=1}^{k} x_{1gc}\theta_{1c} = \sum_{c=1}^{k} x_{1gc} p_c s_{1c}$$

$$E(y_{2h}) = \sum_{c=1}^{k} x_{2hc}\theta_{2c} = \sum_{c=1}^{k} x_{2hc} p_c s_{2c}$$

We denote $\boldsymbol{y}_1 = (y_{11}, \cdots, y_{1m_1})$, $\boldsymbol{y}_2 = (y_{21}, \cdots, y_{2m_2})$, $\boldsymbol{s}_1 = (s_{11}, \cdots, s_{1k})'$ and $\boldsymbol{s}_2 = (s_{21}, \cdots, s_{2k})'$.

$$\boldsymbol{X}_1 = \begin{pmatrix} x_{111} & \cdots & x_{11k} \\ \vdots & \ddots & \vdots \\ x_{1m_11} & \cdots & x_{1m_1k} \end{pmatrix}_{m_1 \times k}$$

$$\boldsymbol{X}_2 = \begin{pmatrix} x_{211} & \cdots & x_{21k} \\ \vdots & \ddots & \vdots \\ x_{2m_21} & \cdots & x_{2m_2k} \end{pmatrix}_{m_2 \times k}$$

$$\boldsymbol{p}_0 = \begin{pmatrix} p_1 & & \\ & \ddots & \\ & & p_k \end{pmatrix}_{k \times k}$$

Step 1: Initialize ctsMultiome from target tissue. We deploy the Joint Non-negative Matrix Factorization model to pre-purify non-RNA profiles, which serves as reference data to generate large-scale pseudo bulk samples used in the training step. We optimize high-dimensional parameters $\boldsymbol{\eta} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{p}_0, \boldsymbol{s}_1, \boldsymbol{s}_2\}$ by minimizing the L_2 loss function:

$$\hat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} \left\{ \|\boldsymbol{y}_1' - \boldsymbol{X}_1 \boldsymbol{p}_0 \boldsymbol{s}_1\|_2^2 + \|\boldsymbol{y}_2' - \boldsymbol{X}_2 \boldsymbol{p}_0 \boldsymbol{s}_2\|_2^2 \right\}$$

subject to $\min\{\boldsymbol{\eta}\} \geq 0$.

This step integrates the observed bulk multiome with shared cell counts to pre-estimate the cell-type-specific profiles in each molecular source, with details in our previous work [35]. The parameters in JNMF are randomly initialized based on the distribution of input bulk multiome and the hypothesized RNA cell counts extracted from previous experiments. This is a reference-free approach that initializes the personalized non-RNA profiles per cell type without using any single cell reference, although the resolution and between-sample variability are limited in the output data.

Step 2: Prepare pseudo-bulk multiome for training. Pseudo-bulk data that imitates the target bulk profile can be generated by aggregating single cell expression data with a given cell type proportion.

We proposed omics-joint pseudo-bulk simulation to connect the different modalities of matched samples by introducing a proportional-shift strategy. The original RNA-seq cell-type proportion could be generated using Dirichlet distribution. Then, a proportion-shift is applied to the original RNA proportion by multiplication with a uniform distribution. Proportional-shift can create the original proportion of non-RNA-seq omics while preserving similarities between the proportions of different molecular sources.

RNA-seq pseudo bulk samples A preprocessed scRNA-seq gene count matrix with cell type labels was used to generate the artificial training samples of RNA-seq. The original RNA-seq cell type proportions were first generated using the Dirichlet() function from the Python package NumPy without *a priori* parameters. Thus, the mean cellular composition across all the training samples are balanced for the cell types of interest. To address the sparsity of training samples heavily biased toward a certain cell type or missing specific cell types, some of the simulated RNA-seq cell type proportions were randomly knocked out and set at zero to build a sparse composition matrix that served as updated ground truth. Next, we determine the exact number of cells to sample from each cell type based on the total number of cells available in the scRNA-seq reference and the ground truth cell type fractions per sample. We then randomly select the appropriate number of cells for each cell type. Finally, the RNA-seq pseudo-bulk sample is produced by aggregating gene expression values across selected single cells.

Non-RNA pseudo bulk samples The generation of training data in the other omics leverages the pre-purified profiles output from JNMF in Step 1. The non-RNA cellular fractions were shifted from the above balanced RNA-seq cell type proportions by multiplying a random shifting factor sampled from the Uniform distribution $U(0.9, 1.1)$ using the NumPy package, randomly scaling the proportions between 90% and 110% of the original value. Sparsity was introduced to this shifted non-RNA proportions through the same knock-out procedure used in RNA-seq pseudo bulk generation. Next, we calculate the mean mixed-cell expression (or abundance) for the input features by the weighted sum of pre-purified non-RNA samples in Step 1, whereas the weights are the cellular composition after random knock-out. To amplify the cross-sample heterogeneity and boost training sample size, a Poisson random generator was applied to the feature-wise mean expression (or abundance) per sample. In this data generating process, we use the pre-purified protein abundance or chromatin accessibility at original (non-log) scale, while the purified DNAm beta values is transformed by negative logarithmic. Since the number of pseudo bulk samples ($N \geq 5000$) is expected to be larger than the sample size of target tissues, the pre-purified profiles in Step 1 can be used repeatedly in this non-RNA pseudo bulk simulation.

Step 3: Multimodal tissue-adaptive dissociation. We propose a multimodal autoencoder to learn the shared molecular signals described by different omics profiles for tissue-adaptive deconvolution. This multimodal autoencoder contains one encoder (E_0) for the concatenated bulk input from RNA and non-RNA profiles and two parallel decoders (D_1, D_2) for the reconstruction of different bulk omics data. The concatenated bulk expression profiles are reduced to latent representation $\boldsymbol{p}_j \in \mathbb{R}^k$ ($j =$

$1, \cdots, N$), representing the ground truth of shared cell counts per training sample, through a five-layer fully connected neural network in molecular modality i ($i = 1, 2$), denoted by $\boldsymbol{W}_i^l$ ($l = 1, \cdots, 5$). We use the encoder to predict cell counts for all the training replicates $\hat{\boldsymbol{P}} = E_0(\boldsymbol{Y})$, whereas $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)$, $\boldsymbol{Y}_1$ is a $N \times m_1$ matrix for the pseudo bulk RNA-seq gene expression profiles and $\boldsymbol{Y}_2$ is a $N \times m_2$ matrix for the simulated non-RNA bulk profiles at training stage. The bulk multi-omic input data are sample-matched with m_1 and m_2 features, respectively. The CELU activation function is used for the first four hidden layers in the encoder, formulated as CELU(x)=max(0,x)+min(0,e^{x-1}). The encoder E_0 maps the high dimensional multi-omic bulk expression profiles to the low dimensional cell count distribution and efficiently captures the nonlinear representations by incorporating non-RNA molecular source. To connect shared encoder with two decoders, two separate fully connected layers without activation function are used to convert the cell count space to the cell proportions for each training replicate:

$$\hat{\boldsymbol{\theta}}_{1j} = \hat{\boldsymbol{p}}_j \cdot \hat{\boldsymbol{S}}_1 \tag{1}$$

$$\hat{\boldsymbol{\theta}}_{2j} = \hat{\boldsymbol{p}}_j \cdot \hat{\boldsymbol{S}}_2 \tag{2}$$

where $\hat{\boldsymbol{\theta}}_{1j}$, $\hat{\boldsymbol{\theta}}_{2j} \in \mathbb{R}^k$ stand for estimated cell proportions of RNA and non-RNA omics; $\boldsymbol{S}_1$, $\boldsymbol{S}_2 \in \mathbb{R}^{k \times k}$ are weight matrices in fully connected layers and can be viewed as modality-specific cell size factors. Two decoders (D_1, D_2) are then leveraged to reconstruct the bulk data of each omics independently based on the cell proportions. Each decoder is composed of five fully connected layers without activation function or bias, which serves as a modality-specific reference panel shared across samples. The decoders are the product of weight matrices ($\boldsymbol{W}_i^l$) with an ReLU function to enforce non-negative values based on its biological meaning. $\hat{\boldsymbol{Y}}_1$ and $\hat{\boldsymbol{Y}}_2$ represent the reconstructed RNA and non-RNA bulk profiles with the same dimension as input data. $\hat{\boldsymbol{X}}_{\mathbf{1}}$ and $\hat{\boldsymbol{X}}_{\mathbf{2}}$ represent the predicted signature matrix of RNA and non-RNA, individually. That is

$$\hat{\boldsymbol{Y}}_1 = D_1(\hat{\boldsymbol{\Theta}}_1) = \hat{\boldsymbol{\Theta}}_1 \cdot \hat{\boldsymbol{X}}_1 = \hat{\boldsymbol{P}} \cdot \hat{\boldsymbol{S}}_1 \cdot \hat{\boldsymbol{X}}_1 \tag{3}$$

$$\hat{\boldsymbol{Y}}_2 = D_2(\hat{\boldsymbol{\Theta}}_2) = \hat{\boldsymbol{\Theta}}_2 \cdot \hat{\boldsymbol{X}}_2 = \hat{\boldsymbol{P}} \cdot \hat{\boldsymbol{S}}_2 \cdot \hat{\boldsymbol{X}}_2 \tag{4}$$

$$\hat{\boldsymbol{X}}_i = ReLU(\boldsymbol{W}_i^1 \cdot \boldsymbol{W}_i^2 \cdot \boldsymbol{W}_i^3 \cdot \boldsymbol{W}_i^4 \cdot \boldsymbol{W}_i^5) \tag{5}$$

We devise the following two-stage procedure to optimize the encoder and decoder parameters per target tissue. In the training stage, we use pseudo-bulk samples of both omics to train the entire MODE framework by minimizing the deconvolution loss L_d and the reconstruction loss L_r. The loss functions are defined as:

$$L_d = \|\boldsymbol{\Theta}_1 - \hat{\boldsymbol{\Theta}}_1\|_1 + \|\boldsymbol{\Theta}_2 - \hat{\boldsymbol{\Theta}}_2\|_1 \tag{6}$$

$$L_r = \|\boldsymbol{Y}_1 - \hat{\boldsymbol{Y}}_1\|_1 + \|\boldsymbol{Y}_2 - \hat{\boldsymbol{Y}}_2\|_1 \tag{7}$$

where $\boldsymbol{\Theta}_i, \boldsymbol{Y}_i$ are the ground truth cell proportions and the bulk multi-omic input for N training samples. The deconvolution loss calculated between the predicted cell proportions and the ground truth is used to optimize the parameters of encoder and neural network connection layers, while the reconstruction loss between the predicted and the input bulk data is used to simultaneously optimize the connection layers for both decoders and encoder. We use L_1 loss in the calculation of deconvolution loss and reconstruction loss.

In the adaptive stage, we employ a two-step optimizing method to train the parameters to adapt to bulk samples. For each target bulk sample, we further optimize the encoder and decoders with tissue-specific loss functions (8) and (9). Let $\tilde{\boldsymbol{\theta}}_{ij'}, \tilde{\boldsymbol{X}}_{ij'}, \tilde{\boldsymbol{y}}_{ij'}$ denote the tissue-adapted cell proportions, high-resolution reference panel, reconstructed bulk profiles for modality i in target tissue j' ($j' = 1, \cdots, n$) based on the parameters learned in the training stage, ensuring the parameters not over fitted compared to the original estimates derived on pseudo bulk training data. The adaptive stage allows personalized deconvolution with parameters re-optimized to each pair of target bulk samples and further improve the prediction beyond training by minimizing the following loss functions.

$$L_{ra} = \|\boldsymbol{y}_{1j'} - \tilde{\boldsymbol{y}}_{1j'}\|_1 + \|\hat{\boldsymbol{X}}_{1j'} - \tilde{\boldsymbol{X}}_{1j'}\|_1 + \|\boldsymbol{y}_{2j'} - \tilde{\boldsymbol{y}}_{2j'}\|_1 + \|\hat{\boldsymbol{X}}_{2j'} - \tilde{\boldsymbol{X}}_{2j'}\|_1 \tag{8}$$

$$L_{da} = \|\boldsymbol{y}_{1j'} - \tilde{\boldsymbol{y}}_{1j'}\|_1 + \|\hat{\boldsymbol{\theta}}_{1j'} - \tilde{\boldsymbol{\theta}}_{1j'}\|_1 + \|\boldsymbol{y}_{2j'} - \tilde{\boldsymbol{y}}_{2j'}\|_1 + \|\hat{\boldsymbol{\theta}}_{2j'} - \tilde{\boldsymbol{\theta}}_{2j'}\|_1 \tag{9}$$

Simulation design for scMultiome in GBM tumors

We designed a rigorous simulation study to mimic the intra-tumor and inter-tumor heterogeneity in bulk multi-omics data. Herein, we used the real single cell multi-omics reference data from an adult GBM tumor tissue profiled by the scRNA-seq and scATAC-seq platforms, individually, available at GEO174554 and https://db.cngb.org/cdcp/dataset/SCDS0000041. The scRNA-seq data were preprocessed and annotated in a previous analysis [36], while the tissue-matched scATAC-seq data still lacked cell labels. Thus, we identified the anchors between RNA and chromatin datasets from the same biological system by using FindTransferAnchor in Seurat 3.0 [13] and then transferred the cell labels of scRNA-seq to scATAC-seq data via function TransferData. The estimate of gene activity matrix for scATAC-seq was quantified by the total counts within 2kb upstream region of the gene body. The annotated real scATAC-seq data along with the tissue-matched scRNA-seq data were visualized in Figure 2a.

To simulate the synthetic bulk multi-omics data, we first randomly sampled 500 cells per modality from five major cellular populations to estimate the cell-specific mean expression and cell-to-cell heterogeneity (Figure 2b). The function fit_marginal in scDesign3 was employed to extract the feature-wise mean and overdispersion, in which the statistical models were zero-inflated Poisson for scRNA-seq counts and Gaussian for scATAC-seq log-scaled counts. The co-expression network was fitted by Gaussian distribution via fit_copula function for each modality. Further, we used the UMAP spatial coordinates extracted by R package uwot from the real scRNA-seq data as the tensor regression covariates along with the above feature-wise parameters in the simulator function (simu_new) to simultaneously generate multiome (gene expression and chromatin accessibility) at cell resolution. This simulator was rerun 200 times to generate a large pool of N=10000 synthetic single cells of multiome with 542 features in RNA and 888 features in ATAC. To generate each pseudo bulk multi-omics sample, we randomly selected 500 synthetic cells without replacement and then aggregated the gene expression or chromatin accessibility across these cells (Figure. 2c). This procedure ensures striking disparities between tumor tissues, since the cells between pseudo bulk samples were distinct.

Declarations

Funding

This work was partially supported by Cancer Center Support Grant P30CA21765 and the American Lebanese Syrian Associated Charities (ALSAC).

Conflict of interest

The authors declared no conflict of interest.

Data availability

- scMultiome tumor simulation based on scRNA-seq and scATAC-seq data in GEO under accession number GEO174554 and https://db.cngb.org/cdcp/dataset/SCDS0000041.
- Human fetal cerebellum organoids: the 10X Genomics scRNA-seq data is available on GEO under accession number GSE150153.
- TCGA adult GBM data: The bulk multi-omics data are available at UCSC Xena and https://gdac.broadinstitute.org/.

Code availability

- MODE (https://github.com/jsuncompubio/MODE)
- scpDeconv (https://github.com/TencentAILabHealthcare/scpDeconv)

- TAPE and Scaden (https://github.com/poseidonchan/TAPE)

Author contribution

This study was conceived by J.S. and Q.L., who developed the algorithm and wrote manuscript. J.S. implemented the algorithm as pipeline, led simulation and real data analyses, summarized and visualized results. W.Z. contributed to deep learning methodology discussion and manuscript writing. Y.P. and Q.L. designed and implemented the simulation experiments. Y.P. and T.L. contributed to real data analysis and visualization. K.S. and P.N. provided the data and interpreted the results in pediatric medulloblastoma. A.O., G.R., P.N. supervised clinical and/or molecular data collection in pediatric medulloblastoma. All authors reviewed the manuscript.

References

[1] Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Yang, T.-H.O., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., *et al.*: The immune landscape of cancer. Immunity **48**(4), 812–830 (2018)

[2] Petralia, F., Ma, W., Yaron, T.M., Caruso, F.P., Tignor, N., Wang, J.M., Charytonowicz, D., Johnson, J.L., Huntsman, E.M., Marino, G.B., et al.: Pan-cancer proteogenomics characterization of tumor immunity. Cell (2024)

[3] Raviram, R., Raman, A., Preissl, S., Ning, J., Wu, S., Koga, T., Zhang, K., Brennan, C.W., Zhu, C., Luebeck, J., *et al.*: Integrated analysis of single-cell chromatin state and transcriptome identified common vulnerability despite glioblastoma heterogeneity. Proceedings of the National Academy of Sciences **120**(20), 2210991120 (2023)

[4] Gray, G.K., Li, C.M.-C., Rosenbluth, J.M., Selfors, L.M., Girnius, N., Lin, J.-R., Schackmann, R.C., Goh, W.L., Moore, K., Shapiro, H.K., *et al.*: A human breast atlas integrating single-cell proteomics and transcriptomics. Developmental cell **57**(11), 1400–1420 (2022)

[5] Wang, L., Jung, J., Babikir, H., Shamardani, K., Jain, S., Feng, X., Gupta, N., Rosi, S., Chang, S., Raleigh, D., *et al.*: A single-cell atlas of glioblastoma evolution under therapy reveals cell-intrinsic and cell-extrinsic therapeutic targets. Nature cancer **3**(12), 1534–1552 (2022)

[6] Russell, A.J., Weir, J.A., Nadaf, N.M., Shabet, M., Kumar, V., Kambhampati, S., Raichur, R., Marrero, G.J., Liu, S., Balderrama, K.S., *et al.*: Slide-tags enables single-nucleus barcoding for multimodal spatial genomics. Nature **625**(7993), 101–109 (2024)

[7] Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y.,

Greenleaf, W.J.: Single-cell chromatin accessibility reveals principles of regulatory variation. Nature **523**(7561), 486–490 (2015)

[8] Cheung, R.K., Utz, P.J.: Cytof—the next generation of cell detection. Nature Reviews Rheumatology **7**(9), 502–503 (2011)

[9] Frangieh, C.J., Melms, J.C., Thakore, P.I., Geiger-Schuller, K.R., Ho, P., Luoma, A.M., Cleary, B., Jerby-Arnon, L., Malu, S., Cuoco, M.S., *et al.*: Multimodal pooled perturb-cite-seq screens in patient models define mechanisms of cancer immune evasion. Nature genetics **53**(3), 332–341 (2021)

[10] Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., *et al.*: Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science **361**(6409), 1380–1385 (2018)

[11] Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., *et al.*: Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nature biotechnology **42**(2), 293–304 (2024)

[12] Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., *et al.*: Integrated analysis of multimodal single-cell data. Cell **184**(13), 3573–3587 (2021)

[13] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., Satija, R.: Comprehensive integration of single-cell data. cell **177**(7), 1888–1902 (2019)

[14] Butler, A., Hoffman, P., Smibert, P., Papalexi, E., Satija, R.: Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature biotechnology **36**(5), 411–420 (2018)

[15] Chen, S., Zhu, B., Huang, S., Hickey, J.W., Lin, K.Z., Snyder, M., Greenleaf, W.J., Nolan, G.P., Zhang, N.R., Ma, Z.: Integration of spatial and single-cell data across modalities with weakly linked features. Nature Biotechnology, 1–11 (2023)

[16] Luo, C., Rivkin, A., Zhou, J., Sandoval, J.P., Kurihara, L., Lucero, J., Castanon, R., Nery, J.R., Pinto-Duarte, A., Bui, B., *et al.*: Robust single-cell dna methylome profiling with snmc-seq2. Nature communications **9**(1), 3824 (2018)

[17] Smith, K.S., Bihannic, L., Gudenas, B.L., Haldipur, P., Tao, R., Gao, Q., Li, Y., Aldinger, K.A., Iskusnykh, I.Y., Chizhikov, V.V., *et al.*: Unified rhombic lip origins of group 3 and group 4

medulloblastoma. Nature **609**(7929), 1012–1020 (2022)

[18] Okonechnikov, K., Joshi, P., Sepp, M., Leiss, K., Sarropoulos, I., Murat, F., Sill, M., Beck, P., Chan, K.C.-H., Korshunov, A., Sah, F., Deng, M.Y., Sturm, D., DeSisto, J., Donson, A.M., Foreman, N.K., Green, A.L., Robinson, G., Orr, B.A., Gao, Q., Darrow, E., Hadley, J.L., Northcott, P.A., Gojo, J., Kawauchi, D., Hovestadt, V., Filbin, M.G., Deimling, A., Zuckermann, M., Pajtler, K.W., Kool, M., Jones, D.T.W., Jäger, N., Kutscher, L.M., Kaessmann, H., Pfister, S.M.: Mapping pediatric brain tumors to their origins in the developing cerebellum. Neuro-Oncology **25**(10), 1895–1909 (2023)

[19] Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., Alizadeh, A.A.: Robust enumeration of cell subsets from tissue expression profiles. Nature methods **12**(5), 453–457 (2015)

[20] Li, B., Severson, E., Pignon, J.-C., Zhao, H., Li, T., Novak, J., Jiang, P., Shen, H., Aster, J.C., Rodig, S., *et al.*: Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome biology **17**, 1–16 (2016)

[21] Wang, X., Park, J., Susztak, K., Zhang, N.R., Li, M.: Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature communications **10**(1), 380 (2019)

[22] Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., Yuan, G.-C.: Accurate estimation of cell-type composition from gene expression data. Nature communications **10**(1), 2975 (2019)

[23] Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., *et al.*: Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature biotechnology **37**(7), 773–782 (2019)

[24] Menden, K., Marouf, M., Oller, S., Dalmia, A., Magruder, D.S., Kloiber, K., Heutink, P., Bonn, S.: Deep learning–based cell composition analysis from tissue expression profiles. Science advances **6**(30), 2619 (2020)

[25] Chen, Y., Wang, Y., Chen, Y., Cheng, Y., Wei, Y., Li, Y., Wang, J., Wei, Y., Chan, T.-F., Li, Y.: Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. Nature Communications **13**(1), 6735 (2022)

[26] Wang, F., Yang, F., Huang, L., Li, W., Song, J., Gasser, R.B., Aebersold, R., Wang, G., Yao, J.: Deep domain adversarial neural network for the deconvolution of cell type mixtures in tissue proteome profiling. Nature Machine Intelligence **5**(11), 1236–1249 (2023)

[27] Bennett, H.M., Stephenson, W., Rose, C.M., Darmanis, S.: Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. Nature Methods **20**(3), 363–374 (2023)

[28] Lee, D.-S., Luo, C., Zhou, J., Chandran, S., Rivkin, A., Bartlett, A., Nery, J.R., Fitzpatrick, C., O'Connor, C., Dixon, J.R., *et al.*: Simultaneous profiling of 3d genome structure and dna methylation in single human cells. Nature methods **16**(10), 999–1006 (2019)

[29] Tian, W., Zhou, J., Bartlett, A., Zeng, Q., Liu, H., Castanon, R.G., Kenworthy, M., Altshul, J., Valadon, C., Aldridge, A., *et al.*: Single-cell dna methylation and 3d genome architecture in the human brain. Science **382**(6667), 5357 (2023)

[30] Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., Kelsey, K.T.: Dna methylation arrays as surrogate measures of cell mixture distribution. BMC bioinformatics **13**, 1–16 (2012)

[31] Drexler, R., Khatri, R., Sauvigny, T., Mohme, M., Maire, C.L., Ryba, A., Zghaibeh, Y., Dührsen, L., Salviano-Silva, A., Lamszus, K., et al.: A prognostic neural epigenetic signature in high-grade glioma. Nature Medicine, 1–14 (2024)

[32] Li, W.V., Li, J.J.: A statistical simulator scdesign for rational scrna-seq experimental design. Bioinformatics **35**(14), 41–50 (2019)

[33] Sun, T., Song, D., Li, W.V., Li, J.J.: scdesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. Genome biology **22**(1), 163 (2021)

[34] Song, D., Wang, Q., Yan, G., Liu, T., Sun, T., Li, J.J.: scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. Nature Biotechnology **42**(2), 247–252 (2024)

[35] Pan, Y., Wang, X., Sun, J., Liu, C., Peng, J., Li, Q.: Multimodal joint deconvolution and integrative signature selection in proteomics. Communications Biology **7**(1), 493 (2024)

[36] Wang, X., Sun, Q., Wang, W., Liu, B., Gu, Y., Chen, L.: Decoding key cell sub-populations and molecular alterations in glioblastoma at recurrence by single-cell analysis. Acta Neuropathologica Communications **11**(1), 125 (2023)

[37] Trobisch, T., Zulji, A., Stevens, N.A., Schwarz, S., Wischnewski, S., Öztürk, M., Perales-Patón, J., Haeussler, M., Saez-Rodriguez, J., Velmeshev, D., *et al.*: Cross-regional homeostatic and reactive glial signatures in multiple sclerosis. Acta neuropathologica **144**(5), 987–1003 (2022)

[38] Sharma, K., Schmitt, S., Bergner, C.G., Tyanova, S., Kannaiyan, N., Manrique-Hoyos, N., Kongi, K., Cantuti, L., Hanisch, U.-K., Philips, M.-A., *et al.*: Cell type–and brain region–resolved mouse brain proteome. Nature neuroscience **18**(12), 1819–1831 (2015)

[39] Jones, D.T., Jäger, N., Kool, M., Zichner, T., Hutter, B., Sultan, M., Cho, Y.-J., Pugh, T.J., Hovestadt, V., Stütz, A.M., *et al.*: Dissecting the genomic complexity underlying medulloblastoma. Nature **488**(7409), 100–105 (2012)

[40] Aldinger, K.A., Thomson, Z., Phelps, I.G., Haldipur, P., Deng, M., Timms, A.E., Hirano, M., Santpere, G., Roco, C., Rosenberg, A.B., *et al.*: Spatial and cell type transcriptional landscape of human cerebellar development. Nature neuroscience **24**(8), 1163–1175 (2021)

[41] Hovestadt, V., Smith, K.S., Bihannic, L., Filbin, M.G., Shaw, M.L., Baumgartner, A., DeWitt, J.C., Groves, A., Mayr, L., Weisman, H.R., *et al.*: Resolving medulloblastoma cellular architecture by single-cell genomics. Nature **572**(7767), 74–79 (2019)

[42] Northcott, P.A., Buchhalter, I., Morrissy, A.S., Hovestadt, V., Weischenfeldt, J., Ehrenberger, T., Gröbner, S., Segura-Wang, M., Zichner, T., Rudneva, V.A., *et al.*: The whole-genome landscape of medulloblastoma subtypes. Nature **547**(7663), 311–317 (2017)

[43] Cedoz, P.-L., Prunello, M., Brennan, K., Gevaert, O.: Methylmix 2.0: an r package for identifying dna methylation genes. Bioinformatics **34**(17), 3044–3046 (2018)

[44] Gibson, P., Tong, Y., Robinson, G., Thompson, M.C., Currle, D.S., Eden, C., Kranenburg, T.A., Hogg, T., Poppleton, H., Martin, J., *et al.*: Subtypes of medulloblastoma have distinct developmental origins. Nature **468**(7327), 1095–1099 (2010)

[45] Feng, Q., Jiang, M., Hannig, J., Marron, J.: Angle-based joint and individual variation explained. Journal of multivariate analysis **166**, 241–265 (2018)

[46] Saulnier, O., Zagozewski, J., Liang, L., Hendrikse, L.D., Layug, P., Gordon, V., Aldinger, K.A., Haldipur, P., Borlase, S., Coudière-Morrison, L., *et al.*: A group 3 medulloblastoma stem cell program is maintained by otx2-mediated alternative splicing. Nature Cell Biology **26**(8), 1233–1246 (2024)

[47] Goldman, M.J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A.N., *et al.*: Visualizing and interpreting cancer genomics data via the xena platform. Nature biotechnology **38**(6), 675–678 (2020)

[48] Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., *et al.*: The cbio cancer genomics portal: an open platform for exploring

multidimensional cancer genomics data. Cancer discovery **2**(5), 401–404 (2012)

[49] Li, Z., Wu, H.: Toast: improving reference-free cell composition estimation by cross-cell type differential analysis. Genome biology **20**, 1–17 (2019)

[50] Baul, S., Tanvir Ahmed, K., Jiang, Q., Wang, G., Li, Q., Yong, J., Zhang, W.: Integrating spatial transcriptomics and bulk rna-seq: predicting gene expression with enhanced resolution through graph attention networks. Briefings in Bioinformatics **25**(4) (2024)

Additional Files

Supplementary File 1 — Supplementary Figures and Tables

Supplementary File 2 — Summary of simulation results