Genome Medicine

https://doi.org/10.1186/s13073-026-01599-w

Article in Press

Multimodal-based analysis of single-cell ATAC-seq data enables highly accurate delineation of clinically relevant tumor cell subpopulations

Received: 28 July 2025

Accepted: 5 January 2026

Published online: 13 January 2026

Cite this article as: Xiong K., Wang W., Ding R. *et al.* Multimodal-based analysis of single-cell ATAC-seq data enables highly accurate delineation of clinically relevant tumor cell subpopulations. *Genome Med* (2026). https://doi.org/10.1186/s13073-026-01599-w

Kewei Xiong, Wei Wang, Ruofan Ding, Dinglin Luo, Yangmei Qin, Xudong Zou, Jiguang Wang, Chen Yu & Lei Li

Multimodal-based analysis of single-cell ATAC-seq data enables highly accurate delineation of clinically relevant tumor cell subpopulations

Kewei Xiong[1,4], Wei Wang[2,4], Ruofan Ding[1], Dinglin Luo[1], Yangmei Qin[1], Xudong Zou[1], Jiguang Wang[3], Chen Yu[2*], Lei Li[1*]

[1]Institute of Systems and Physical Biology, Shenzhen Bay Laboratory; Shenzhen 518055, China.

[2]Institute of Cancer Research, Shenzhen Bay Laboratory; Shenzhen 518055, China.

[3]Division of Life Science, Department of Chemical and Biological Engineering, State Key Laboratory of Molecular Neuroscience, The Hong Kong University of Science and Technology; Hong Kong SAR, China

[4]These authors contributed equally to this work.

[*]Corresponding author. Email: yu@szbl.ac.cn; lei.li@szbl.ac.cn

Abstract

Background

Accurately identifying functionally distinct tumor cell subpopulations remains a critical challenge in cancer research. While single-cell epigenomics assays provide powerful insights into tumor heterogeneity beyond gene expression, computational limitations have hindered their application.

Methods

We introduce Multimodal-based Analysis of scATAC-Seq data (MAAS), a method that integrates chromatin accessibility, copy number variations (CNVs), and single-nucleotide variants (SNVs) to identify functional tumor cell subpopulations. MAAS employs a self-expressive multimodal matrix factorization approach with rigorous coverage normalization and data denoising. We applied MAAS to simulated datasets and multiple real-world tumor scATAC-seq datasets, including pediatric ependymoma, B-cell lymphoma, and glioblastoma, and benchmarked its performance against existing integration methods. Functional relevance of subpopulation-specific genes was experimentally validated using gene knockdown and overexpression assays. Furthermore, we constructed subpopulation-specific gene regulatory networks and developed a prognostic signature from the key regulatory genes.

Results

MAAS demonstrated superior accuracy in detecting clinically relevant subpopulations, particularly in tumors with limited CNV heterogeneity, such as pediatric ependymoma and B-cell lymphoma. In glioblastoma, MAAS uncovered a previously unrecognized subpopulation with temozolomide resistance and further experimentally validated the effects of its signature genes

through gene knockdown and overexpression. The MAAS-derived prognostic signature, MAASig, outperformed traditional clinicopathologic features across multiple cancer types when applied to independent validation cohorts.

Conclusions

By integrating multimodal information from scATAC-seq data, MAAS provides the robust identification of functionally distinct tumor cell subpopulations, facilitating the discovery of potential therapeutic targets.

Keywords: scATAC; multimodal analysis; tumor heterogeneity; drug resistance; prognostic signature

Background

Cancer cells undergo various genetic and epigenetic changes that drive the formation of distinct subpopulations during tumor progression [1, 2]. Identifying these critical subpopulations accurately is essential for developing effective treatments [2]. Single-cell sequencing technologies have revolutionized our understanding of tumor cellular composition compared to traditional bulk analyses. Among these, single-cell RNA sequencing (scRNA-seq) is widely used to profile clinical phenotype-associated subpopulations [3-5]. Although scRNA-seq can distinguish malignant from non-malignant subpopulations and examine cell heterogeneity by analyzing expression-derived genetic mutations such as copy number variations [6, 7], it often fails to identify the clinically relevant subpopulations that are not solely linked to gene expression [8]. This limitation arises because scRNA-seq primarily captures transcriptional activity and often overlooks critical epigenetic and regulatory changes [9, 10].

Single-cell epigenomics assays, such as single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq), enable robust profiling of cell subpopulations beyond gene expression and can reveal distinct regulatory information that controls gene expression. However, applying scATAC-seq to study tumor subpopulations has been computationally challenging due to inherent data sparsity, technical noise, and variable cell-to-cell sequencing depth that can confound clustering analyses. Traditional methods, such as Copy-scAT and epiAneufinder, mainly rely on CNVs [11-13], to determine the genetic heterogeneity of tumor cells. Despite these advances, CNV analysis alone often fails to detect clinically relevant tumor subpopulations. For example, melanoma subpopulations with varying anti-PD-1 responses are better characterized by their distinct point mutation profiles rather than CNV events [7]. Furthermore, the interplay between genetic and epigenetic changes enables subpopulations to

circumvent therapeutic barriers, promoting cancer progression [14, 15], thus highlighting the need to integrate these features. Unfortunately, existing methods fail to fully utilize the spectrum of epigenetic and genetic information available from scATAC-seq data. These methods are also limited in delineating tumor cell subpopulations with low CNV heterogeneity, such as those found in hematopoietic and pediatric cancers [16]. Additionally, scATAC-seq data often exhibit inherent high sparsity and technical noise, posing significant challenges in determining informative features for dissecting tumor cell subpopulations [17].

Here, we presented a novel computational method called Multimodal-based Analysis of scATAC-Seq data (MAAS) that accurately identifies tumor cell subpopulations and infers their evolutionary lineages by integrating informative multimodal features, including CNVs, single-nucleotide variants (SNVs), and chromatin accessibility data. To overcome the technical challenges of scATAC-seq analysis, MAAS implemented rigorous normalization procedures to correct for variable cell coverage and employed robust denoising strategies for sparse SNV data. Our approach quantitatively assesses the contribution of each data modality, allowing for the identification of subpopulations driven predominantly by one modality, particularly in tumors with limited CNV heterogeneity. MAAS outperformed other state-of-the-art methods on both simulated and real datasets. In pediatric ependymoma, a cancer with low CNV heterogeneity, MAAS identified a progressive tumor cell subpopulation associated with multidrug resistance. When applied to a glioma tumor, MAAS uncovered a previously overlooked subpopulation resistant to temozolomide, which was subsequently experimentally validated. Furthermore, we developed a MAAS-derived multimodal clinical signature by integrating subpopulation-specific gene regulatory networks (GRNs), which provided a more accurate prognostic prediction than traditional clinicopathologic characteristics and existing signatures across multiple cancer types.

In conclusion, MAAS is a reliable and robust tool for identifying clinically relevant tumor cell subpopulations, facilitating the discovery of new disease mechanisms and enhances tumor diagnosis and therapeutic strategies.

Methods

Datasets used in this study

All the data in this study are publicly available. The scATAC-seq datasets used for MAAS analysis comprised the K562 cell line (n = 2 from GSE243430) [18], the SNU601 gastric cancer cell (n = 1 from PRJNA674903) [19], two ovarian cancer (OC) cohorts (n = 3 from phs002340.v1.p1 [20] and n = 10 from GSE247982 respectively [21]), pediatric posterior fossa ependymoma (PPFE) (n = 4 from GSE206579) [22], B-cell lymphoma (n = 1) [23], adult glioblastoma (GBM) (n = 4 from GSE139136) [24], pediatric GBM (n = 3 from GSE163655) [24], hepatocellular carcinoma (HCC) (n = 13 from GSE227265) [25], clear cell renal cell carcinoma (ccRCC) (n = 19 from GSE207493) [26]. In addition, an scATAC-seq dataset coupled with whole-exome sequencing data (n = 2 from PRJNA533341) was used for validation of CNV calling [27].

Bulk RNA-seq datasets used for drug resistance analysis included PPFE (GSE42658, n = 14; [28]; GSE13267, n = 17 [29]; GSE66354, n = 55 [30]) and GBM (GSE53014, n = 12 [31]; GSE68029, n = 12 [32]; CGGA693, n = 289; and CGGA325, n = 139 [33]). Bulk ATAC-seq data for B-cell lymphoma with treatment information were obtained from GSE254913 (n = 8) [34].

Bulk RNA-seq datasets for signature analysis included GBM (The Cancer Genome Atlas (TCGA), n = 153 [35]; CGGA [33]), OC (TCGA, n = 357 [35]; GSE140082, n = 380 [36]; GSE32062, n = 270 [37]), B-cell lymphoma (GSE181063, n = 1310 [38]; GSE10846, n = 420 [39]; GSE136971, n = 448 [40]), hepatocellular carcinoma (TCGA, n = 341 [35]; GSE116174, n = 64 [41]; GSE76427, n = 115 [42]), clear cell renal cell carcinoma (TCGA, n = 504 [35]; E-MTAB-1980, n = 92 [43]; CPTAC, n = 53 [44]). In all datasets, n denotes the number of samples.

scATAC-seq data analysis

We used the SRA Toolkit (v2.10.9) [45] to obtain FASTQ files of raw sequencing data, which were then aligned to the GRCh38 reference genome using 10× Genomics Cell Ranger ATAC (v2.1.0) software [46] with default parameters. We then used the Signac [47] package to obtain a cell-by-peak matrix. High-quality cells were retained based on transcription start site enrichment (> 3), the number of unique fragments (> 1000), percentage of reads in peaks (> 15%), blacklist ratio (< 5%) and nucleosome signal (< 4). To account for varying coverage across cells, we performed term frequency-inverse document frequency (TF-IDF) normalization, applying a log-transformation to both the TF and IDF elements [48]:

$$Normalized\ \ peak = \log(TF) \times \log(IDF) = \log\left(\frac{C_{ij}}{F_j}\right) \times \log\left(\frac{N}{n_i}\right) \quad (1)$$

where C_{ij} is the total number of counts for peak i in cell j and F_j is the total number of counts for cell j. For the IDF term, N denotes the total number of cells and n_i represents the total number of counts for peak i across all cells. To correct batch effects across samples, we performed ComBat [49] analysis using the R package sva [50]. In addition, a cell-by-gene score matrix used for functional enrichment analysis was obtained using the GeneActivity function implemented in

Signac. Differentially accessible chromatin regions (DACRs) were identified using the FindAllMarkers function by regressing out the library size. Details of tumor cell identification were provided in the Additional file 1: Supplementary Methods. The quality control metrics of each dataset analyzed in this study were summarized in Additional file 1: Table S1.

Mutation calling

We benchmarked two CNV callers (epiAneufinder [12] and Copy-scAT [11]) for scATAC-seq data. Based on the benchmarking results, we selected epiAneufinder [12] for MAAS analysis (Additional file 1: Figs. S1 and S2). Additionally, we employed SComatic [51] for somatic SNV calling and evaluated seven tools for SNV denoising (Additional file 1: Figs. S3-S6). Based on the results from the SNV denoising benchmark, we selected CBM [52] as the preferred method for this study. Further details are provided in Additional file 1: Supplementary Notes 1-2 and Methods.

Cell affinity estimation in each feature layer

We first corrected the chromatin accessibility profile according to the prior knowledge that copy number gain leads to an aberrant high peak density and vice versa.

$$\widetilde{x}_{pj}=\begin{cases} x_{pj}-\xi\times\left(r_{p\varepsilon(j)}-2\right)\times x_{pj}, & if \quad j\cap R^{+}_{p\varepsilon(j)}\neq\varnothing \ \text{ and } \ x_{pj}\geq 2; \\ x_{pj}+\xi\times\left(2-r_{p\varepsilon(j)}\right)\times x_{pj}, & if \quad j\cap R^{-}_{p\varepsilon(j)}\neq\varnothing \ \text{ and } \ x_{pj}>0; \\ x_{pj}, & otherwise \end{cases} \tag{2}$$

where x_{pj} and $\widetilde{x}_{pj}$ indicate the raw and adjusted peaks j of cell p, respectively, and $r_{p\varepsilon(j)}$ indicates the observed copy numbers containing the region j. The hyperparameter ξ indicates the prior regarding the effect of copy number on chromatin accessibility, with values ranging from 0

to 1 (default: 0.5). We demonstrated the necessity of this correction, as well as its robust performance across different values of ξ (Additional file 1: Figs. S7-S9). A higher value of ξ indicates a stronger assumed influence of copy number on chromatin accessibility levels. Specifically, when ξ is set to 1, it suggests that increased chromatin accessibility is entirely dependent on copy numbers. Conversely, lower values of ξ reflect a weaker or more nuanced relationship between CNVs and chromatin accessibility. This parameter facilitates flexible modeling of the extent to which CNVs are presumed to drive changes in chromatin accessibility. $R^{+}_{p\varepsilon(j)}$ and $R^{-}_{p\varepsilon(j)}$ represent the copy number gain and loss region of cell p, respectively. Then, we calculated the affinity between cell p and q (default: cosine). The Hamming distance was used to estimate the cell similarity based on CNVs or SNVs.

MAAS structure

We employed a modified non-negative matrix factorization to jointly integrate multiple modalities for dimension reduction of affinity matrices $A^{(i)}$. The model aimed to identify a consensus low-dimensional space W that simultaneously encodes different layers, along with diagonal matrices $H^{(i)}$ to representing the coefficients of latent factors to be projected into this space:

$$A^{(i)} \sim A^{(i)} W H^{(i)} W^{T} \quad (3)$$

We noted that $A^{(i)}$ serve as self-expressive terms, indicating that our multimodal integration method can learn and maintain local structure for subspace clustering (Additional file 1: Supplementary Note 3). Given the input terms, our model minimizes the loss function as follows:

$$Q(W,\ H^{(i)};\ A^{(i)})=\frac{1}{2}\sum_{i}\left\|A^{(i)}-A^{(i)}W H^{(i)}W^{T}\right\|_{F}^{2} \tag{4}$$

Multiplicative update rules were utilized through the stochastic gradient descent as follows:

$$\begin{aligned} W &\leftarrow W-\varphi\ \nabla_{W}Q(A^{(i)},\ W,\ H^{(i)}) \\ H^{(i)} &\leftarrow H^{(i)}-\eta\ \nabla_{H^{(i)}}Q(A^{(i)},\ W,\ H^{(i)}) \end{aligned} \tag{5}$$

where

$$\begin{aligned} \nabla_{W}Q &=-\sum_{i}\left[A^{(i)}(A^{(i)}-A^{(i)}W H^{(i)}W^{T})W H^{(i)}+(A^{(i)}-W H^{(i)}W^{T}A^{(i)})A^{(i)}W H^{(i)}\right] \\ \nabla_{H^{(i)}}Q &=-[Z_{0}^{(i)}]^{T}(A^{(i)}-Z_{0}^{(i)}H^{(i)}W^{T})W,\ \ where\ \ Z_{0}^{(i)}=A^{(i)}W \end{aligned} \tag{6}$$

Therefore, we could obtain that

$$\begin{aligned} W-\varphi\ \nabla_{W}Q(A^{(i)},\ W,\ H^{(i)}) &= W+2\varphi\sum_{i}[A^{(i)}]^{2}W H^{(i)}-\varphi\sum_{i}[A^{(i)}(A^{(i)}W H^{(i)}W^{T})+(\\ H^{(i)}-\eta\ \nabla_{H^{(i)}}Q(A^{(i)},\ W,\ H^{(i)}) &= H^{(i)}+\eta[Z_{0}^{(i)}]^{T}W-\eta[Z_{0}^{(i)}]^{T}(Z_{0}^{(i)}H^{(i)}W^{T})W \end{aligned} \tag{7}$$

Based on the derivatives, the learning rate for the rule was denoted as

$$\begin{aligned} \varphi &= \frac{W}{\sum_{i}[A^{(i)}(A^{(i)}W H^{(i)}W^{T})+(W H^{(i)}W^{T}A^{(i)})A^{(i)}]W H^{(i)}} \\ \eta &= \frac{H^{(i)}}{[Z_{0}^{(i)}]^{T}(Z_{0}^{(i)}H^{(i)}W^{T})W} \end{aligned} \tag{8}$$

A block coordinate descent scheme was implemented, in which we optimized based on only one rule and kept others fixed. Finally, we implemented the decompositions using hand-solved equations

$$\begin{aligned} W &\leftarrow 2\times W\quad \frac{\sum_{i}[A^{(i)}]^{2}W H^{(i)}}{\sum_{i}[A^{(i)}(A^{(i)}W H^{(i)}W^{T})+(W H^{(i)}W^{T}A^{(i)})A^{(i)}]W H^{(i)}} \\ H^{(i)} &\leftarrow H^{(i)}\quad \frac{[Z_{0}^{(i)}]^{T}A^{(i)}W}{[Z_{0}^{(i)}]^{T}(Z_{0}^{(i)}H^{(i)}W^{T})W} \end{aligned} \tag{9}$$

The gradient descent terminated when the condition $\frac{Q_i}{Q_{i-1}-Q_i} < 10^{-6}$ could be met (Additional file 1: Fig. S10). The optimized process of our model is provided in Additional file 1: Supplementary Note 4. Tumor cell subpopulations were identified by applying *K*-means clustering to the latent factor matrix **W**. To determine the optimal number of clusters, we systematically evaluated a range of cluster numbers (from $k = 2$ to $k = 10$) and the dimensions of **W** (from 2 to 7), and calculated four widely used clustering validity indices [53-56]: the silhouette index, Davies-Bouldin index, Dunn validity index, and Calinski-Harabasz index. These metrics respectively assess intra-cluster cohesion, inter-cluster separation, cluster compactness, and overall partition quality. To integrate these into a single measurement, we proposed a composite score termed the *S*-score (Additional file 1: Supplementary Methods), which normalizes and combines the four indices into a weighted sum. The cluster number with the highest *S*-score was selected as the optimal resolution, and the corresponding *K*-means partition was used to define tumor cell subpopulations. The *S*-scores of each clustering assignment based on MAAS embedding across datasets were summarized in Additional file 1: Fig. S11.

Contribution of modalities

We quantified the contribution of each modality by calculating the normalized trace, defined as the sum of diagonal values of the corresponding matrix. The contribution of modality *i* is given by

$$Contribution\ of\ modality\ i = \frac{trace\left(H^{(i)}\right)}{\sum_t trace\left(H^{(t)}\right)} = \frac{\left\|H^{(i)}\right\|_1}{\sum_t \left\|H^{(t)}\right\|_1} \quad (10)$$

where $trace(\cdot)$ represents the sum of the diagonal elements, and $\|\cdot\|_1$ denotes the L1 norm. A larger trace value indicates a higher weight or greater influence of the corresponding modality in cluster assignment. Namely, this metric allows for a quantitative comparison of the relative importance of different modalities in predicting tumor subpopulations.

Construction of cell hierarchy

The consensus cell-affinity matrix was initially computed based on W. Alternatively, it could be derived from the features of individual modalities. The evolution tree was reconstructed using the minimum evolution algorithm [57], as implemented in the R package ape [58]. To visualize the resulting tree, we employed the ggtree [59] package, utilizing the "ape" layout.

Tumor cell identification

We employed a comprehensive approach to identify tumor cells based on multiple criteria. For each dataset, tumor cells were identified using a combination of: 1) CNV profiles characteristic of the cancer type, 2) marker gene accessibility specific to cancer cells, 3) known tumor-specific epigenetic signatures, and 4) clustering patterns consistent with malignant populations. For the glioma dataset specifically, tumor cells were identified based on characteristic CNVs including chromosome 7 gain and chromosome 10 loss, which are hallmark alterations in glioblastoma.

Multidrug sensitivity of PPFE cell subpopulations

We used scRank [60] to calculate perturbation scores of each drug, including etoposide, vinblastine and vincristine. Edges with weight lower than 0.9 were removed. To identify gene modules across samples, we used the NMF implemented in the R package GeneNMF [61] by

setting the number target NMF components for each sample from 4 to 9, and 10 target meta-modules were determined by hierarchical clustering with minimum confidence of 0.1. We then estimated drug-target module activity using the AddModuleScore function in the Seurat package [62]. Additionally, we used oncoPredict [63] to calculate the half maximal inhibitory concentration (IC_{50}) of ependymoma patients from the GSE13267 [29] and GSE66354 [64] cohorts.

TMZ response of GBM tumor cell subpopulations

We first used oncoPredict [63] to predict IC_{50} of GBM patients in the CGGA693, CGGA325 and TCGA cohorts, respectively, by performing linear regression. Specifically, drug response data from GDSC2 [65] were utilized as the training set, while the three bulk RNA-seq datasets served as the test set. Genes with a median absolute deviation less than 0.15 were excluded from the regression analysis. We then used the calcPhenotype function to predict IC_{50} for each patient. The parameters were set as follows: powerTransformPhenotype set to FALSE, removeLowVaryGenes set to 0.2, removeLowVaringGenesFrom specified as 'rawData', and minNumSamples set to 10. Patients were subsequently stratified into sensitive and resistant groups based on the median cutoff of IC_{50} score. Next, we applied Scissor [3] to predict the therapeutic phenotype of each cell using binomial regression model. Prediction performance was estimated using the reliability.test function with 1000 permutation times and 10-fold cross-validation. Additionally, we calculated the E-distance between tumor cell clusters and experimentally determined TMZ-sensitive and resistant subpopulations using the edist function from the R package Rfast [66], based on gene activity inferred from scATAC-seq data. To reduce the impact of sample

size on distance computation, we randomly selected 100 cells from each cluster and repeated this procedure 500 times.

Identification of cluster 2-specific genes for experimental validation

We screened cluster 2-specific genes for experimental validation of TMZ resistance based on the following steps: Firstly, we identified genes with significant increasing expression in TMZ-resistant GBM cell lines [67] by GEO2R [68], including LNZ308 and U251, with the thresholds of logFC > 0.25 and adjusted *P*-value < 0.1. The IC_{50} values of the resistant GBM subpopulations showed > 2-fold increase in TMZ-resistance compared to the parental cell lines [67]. Then, we selected genes with prognostic significance by both log-rank test and univariate Cox regression with the thresholds of HR > 1 and *P*-value < 0.05. Finally, we examined the overlapped TMZ-resistance relevant genes, survival-related genes and cluster 2-specific genes.

Cell culture

HEK293T, U-87 MG cells (Cell bank of Chinese Academy of Sciences, Shanghai), and U-251 MG cells (a generous gift from Dr. Tengfei Guo) were cultured in high-glucose DMEM containing 10% FBS (VISTECH) and 1% penicillin/streptomycin (Thermo Fisher Scientific). All cell lines were routinely tested and confirmed to be free of mycoplasma contamination was detected during cell culture.

RNA preparation and quantitative RT-PCR

Total RNA was extracted with Quick-RNA™ Miniprep Kit (Zymo Research), followed by cDNA

synthesis with a GoScript™ Reverse Transcription System (Promega). Quantitative RT-qPCR was performed in a Real-Time PCR system (Bio-Rad) using SYBR Green Supermix (CWBIO). The primer sequences used are listed in Additional file 1: Table S2.

Lentivirus preparation and titration

To construct lentiviral vectors expressing shRNA targeting *TPST1*, *RFTN1* and *ADAMTS1*, corresponding shRNA oligonucleotides (Additional file 1: Table S3) were inserted into the cloning site of pLKO.1 (Addgene, # 10878) following the manufacturer's instructions. All constructs were validated by Sanger sequencing. Lentivirus was packaged as previously described [69]. Briefly, viruses were harvested from HEK293T cells transfected with the indicated plasmid and the packaging plasmids pMD2G and psPAX2 using PEI MAX transfection reagents (Polysciences), concentrated and titrated. For virus titration, viruses were tested by counting the U-87 MG or U251 MG cell clones after 3 μg/mL puromycin (Beyotime Biotechnology) selection.

Cell proliferation assay

For cell proliferation assay, U-87 MG cells were infected with indicated shRNA targeting *TPST1*, *RFTN1*, and *ADAMTS1*. After 3 μg/mL puromycin selection, U-87 MG cells were re-plated 1000 cells per well on 96-well plates. Cell viability was determined using Cell Counting Kit-8 assays (CCK-8) by measuring the absorbance at 450 nm using a microplate reader (BioTek), following the manufacturer's instructions (Beyotime Biotechnology).

TMZ chemosensitivity assay

As previously described [70]{Zou, 2021 #433}, U-87 MG and U251 MG cells were infected with indicated shRNA or *TPST1* overexpression vector, after 3 μg/mL puromycin selection. Subsequently, the surviving cells were allowed to recover in fresh growth medium for 2 days. The infected cells were re-plated 5000 cells per well on 96-well plates. Following 24 hours of incubation, the medium was replaced with fresh medium containing 50 μM or 200 μM TMZ. Cells were then cultured for another 24 hours and cell viability was assessed using the CCK-8 (Beyotime Biotechnology) according to the manufacturer's protocol. Absorbance at 450 nm was measured using a BioTek microplate reader to determine viability.

Statistical analysis

The Wilcoxon rank-sum test [71] or Student's t-test [72] (sample size < 10) were used to compare quantitative measures between groups of interest. Comparisons of relative frequencies were performed by Fisher's exact test [73]. In addition, we performed several survival analyses to investigate the prognostic relevance of tumor subpopulation-specific genes. Samples were stratified into two groups based on the median cutoff. Survival curves of the two patient groups were evaluated using the Kaplan-Meier approach [74]. The statistical significance was calculated using a two-tailed log-rank test. We used the survival R package [75] for Cox analysis and the two-tailed Wald test [76]. Time-dependent AUC and C-index were calculated using the R packages survivalROC [77] and survival [75], respectively.

Results

MAAS achieved superior accuracy in predicting tumor cell subpopulations

To delineate cellular heterogeneity in tumors, we developed an algorithm called MAAS to accurately identify tumor cell subpopulations by integrating genetic and epigenetic features derived from scATAC-seq data (Fig. 1a). Specifically, MAAS infers a consensus low-dimensional latent factor that encodes multiple modality features, including CNVs, SNVs and chromatin accessibility, which are subsequently used for tumor cell subpopulation identification. To mitigate potential biases from CNVs, which often confound the quantification of chromatin accessibility [78], MAAS incorporates a weighted correction strategy to adjust for this effect. Additionally, given the sparsity and noise inherent in SNVs derived from scATAC-seq data, MAAS utilized a parametric algorithm CBM [52], which outperforms other denoising algorithms in correcting false discoveries according to our benchmarking analysis, thus enabling accurate profiling of somatic mutations in individual cells (Fig. 1a; Additional file 1: Figs. S4-S6 and Supplementary Note 2). Cell similarities were then estimated using cosine distance for chromatin accessibility and hamming distance for CNV and SNV. MAAS integrated these three cell-by-cell matrices $A^{(i)}$ (i = 1, 2, 3) using multimodal non-negative matrix factorization, optimized by a multiplication update algorithm, generating a latent variable $\mathbf{W}$ and three diagonal coefficient matrices $H^{(i)}$ upon convergence (Fig. 1b). Notably, MAAS leverages correlation matrices $A^{(i)}$ as self-expressions to further enhance clustering accuracy [79], and the contribution of each modality is estimated by the trace of $H^{(i)}$ (Fig. 1c). Finally, tumor cells are classified into subpopulations using *K*-means clustering, and a minimum evolution tree depicting subpopulation relationships is constructed (Fig. 1c).

To systematically evaluate the performance of MAAS, we conducted a simulation analysis to determine its ability to accurately deconvolute labeled tumor cell subpopulations. We first generated three simulated cell clusters as ground truth datasets, where clusters 1 and 2 shared

different genetic features compared to cluster 3, while also exhibiting distinct chromatin accessibility profiles (Fig. 2a). MAAS successfully separated clusters 1 and 2, which were indistinguishable using single-modality approaches (Fig. 2a and b). Specifically, MAAS identified 92.71%, 94.85%, and 95.12% of the three clusters, respectively (Fig. 2c). When tested on four simulated cell clusters, MAAS accurately identified 93.96%, 95.17%, 98.48% and 56.03% of cells in clusters 1 to 4, respectively (Additional file 1: Fig. S12a-c).

To further evaluate the robustness of the MAAS method, we compared it with other multi-omics integration and clustering tools, including uniform manifold approximation and projection (UMAP) [80], intNMF [81], PintNMF [82], SNF [83], LRACluster [84], MCIA [85], MOFA [86], multiVI [87], MOJITOO [88], SEACells [89], scOpen [90] and CoGAPS [91] (Additional file 1: Supplementary Methods). We began by randomizing the tumor cell clusters and evaluated the performance using three metrics: the adjusted Rand index (ARI), normalized mutual information (NMI), and V-measure. MAAS significantly outperformed UMAP multimodal clustering (Additional file 1: Fig. S12d), other integration methods, and the single-modality method CBM [52] with the median ARI, NMI, and V-measure values of 0.912, 0.833, and 0.849, respectively (Fig. 2d). Additionally, we varied the number of cells per tumor cluster for each simulation and found that MAAS consistently achieved the highest ARI, NMI, and V-measure scores, and cell number had a minimal effect on MAAS performance (Additional file 1: Figs. S12e and S13). Further examining the impact of the cluster number on clustering performance demonstrated that MAAS outperformed alternative methods across a diverse range of cluster sizes (Additional file 1: Fig. S12e and f). We also evaluated clustering performance under different levels of data sparsity, ranging from 10% to 90%, and found MAAS consistently showed superior performance, maintaining over 40% correct classification even at the 90% of data sparsity

(Additional file 1: Fig. S12g). To validate the accuracy of cell hierarchy reconstruction, we utilized mutual cluster information, a generalized Robinson-Foulds metric [92]. Notably, MAAS demonstrated superior performance across a range of subpopulation numbers, achieving a median mutual cluster information score of 0.811. In contrast, MOFA and intNMF showed markedly lower performance, with median scores of 0.719 and 0.595, respectively (Additional file 1: Fig. S14). Ablation analysis revealed that the simultaneous integration of chromatin accessibility, CNVs, and SNVs outperformed any pairwise combination or single-modality approach (Additional file 1: Fig. S15). Additionally, we evaluated the computational efficiency with respect to the total number of cells. We found that MAAS was more scalable than other matrix factorization-based methods such as CoGAPS and PintMF, ranging from ~1 hours and 0.4GB for 400 cells to ~44 hours and 24 GB random access memory for 20,000 cells (Additional file 1: Fig. S16).

Moreover, we applied MAAS to a K562 dataset which generates both ATAC and whole-genome sequencing (WGS) data from the same cell [93]. The overall clustering results showed high consistency with those obtained using scATAC-seq data alone (Additional file 1: Fig. S17). We also compared the performance of MAAS with CNV estimates derived from single-cell WGS (scWGS) data for gastric cancer [13]. MAAS accurately recovered the four tumor cell clusters characterized by the amplification of chromosomes 1 and 3 and the deletion of chromosomes 4 and 18, with an average Pearson's correlation of 0.885 (Additional file 1: Fig. S18). To further access the performance of MAAS in identifying clinically pertinent tumor cell subpopulations, we benchmarked MAAS in a real OC scATAC-seq dataset from three tumors [94], the MAAS method effectively identified metastatic tumor cells from primary ones and accurately distinguished tumor cells at different pathological stages (Fig.

2g and h and Additional file 1: Fig. S19). In another OC dataset [95], MAAS accurately distinguished treated from non-treated cells, and revealed the heterogeneity within both treated and non-treated populations by identifying seven subpopulations characterized by distinct cluster-specific CNVs, such as gain of chromosomes 1q (C5), 12p (C6) and 19q (C1), as well as losses of chromosomes 8p (C6) and 9p (C5) (Additional file 1: Fig. S20). Overall, MAAS demonstrated superior performance in predicting tumor cell subpopulations compared to state-of-the-art methods.

MAAS detected clinically relevant tumor cell subpopulations with low CNVs

Many tumors, such as pediatric ependymoma, exhibit a low frequency of CNV events, often less than 10% [96], posing a great challenge for traditional methods in resolving tumor heterogeneity. To demonstrate the utility of the MAAS method in detecting subpopulations with low CNVs, we applied it to a scATAC-seq dataset of PPFE [97]. Our analysis revealed that MAAS accurately predicted the three major tumor cell subpopulations from 2,428 tumor cells (Additional file 1: Figs. S21 and S22), providing a clearer distinction between tumor cell clusters than the traditional methods (Fig. 3a-d and Additional file 1: Fig. S23). Functional enrichment analysis of the MAAS-predicted clusters showed that the MAAS not only recovered traditional hallmark cancer pathways but also identified an additional subset of cancer-related pathways enriched specifically in clusters 1 and 3, such as DNA repair, *E2F* targets and *p53* pathway (Fig. 3b). This suggests clusters 1 and 3, driven primarily by chromatin accessibility and SNVs, represents functional subpopulations detected by MAAS (Fig. 3c). We then evaluated the proliferation and migration characteristics of each cluster using key proliferation signature genes [98, 99] (*MKI67*, *PCNA*, *IGF1*, *ITGB2*, *PDGFC*, *JAG1*, *PHGDH*, *BCL2*) and migration-related genes [100, 101]

(*ARID5B* and *FAT1*) (Additional file 1: Supplementary Methods). Cluster 1 exhibited significantly higher proliferation and migration scores than clusters 2 and 3 (Fig. 3d), indicating its strong metastatic potential and highly aggressive phenotype. Additionally, deconvolution analysis based on cluster-specific genes applied to a bulk RNA-seq pediatric ependymoma dataset [102] revealed that patients with grade III tumors exhibited significantly higher abundance of cluster 1 compared to those with grade II tumors (Fig. 3e and Additional file 1: Supplementary Methods). Given the highly differential chromatin accessibility profiles, we reasoned that these differences might reflect dynamic cellular state changes. Therefore, we used Monocle [103] to reconstruct developmental trajectories and observed two stepwise transitions: from cluster 2 to cluster 1 and cluster 2 to cluster 3 (Fig. 3g and Additional file 1: Supplementary Methods). To further validate the dynamic changes between MAAS-predicted clusters, we generated a minimum-evolution tree to depict the evolutionary process and found that cluster 3 had the highest mutation burden and chromatin accessibility, followed by cluster 1 (Fig. 3h and i; Additional file 1: Figs. S24-S25).

We then investigated the response of MAAS-determined clusters to several first-line chemotherapeutics, including etoposide, vinblastine and vincristine. First, we estimated the drug-resistant subpopulation using scRank [104] (Methods) and observed that new cluster 1 exhibited lower perturbation scores than clusters 2 and 3 for each of the three drugs (Fig. 3j), suggesting that cluster 1 is more drug-tolerant. Additionally, cluster 1 had significantly lower scores for the drug-target gene modules of all three drugs, as determined by the activity of the drug-target gene module (Fig. 3k and l; Additional file 2: Data S1; Wilcoxon rank-sum test, all P-values < 0.05). We then performed deconvolution analysis for pediatric ependymoma patients (Supplementary Methods) and found that drug-resistant samples contained an average of 85.71% more cluster 1 cells

(Fig. 3m). To identify potential targeted therapies for cluster 1, we screened data from the LINCS consortium [105] for compounds that selectively target cluster 1-specific transcription factors and kinases (Additional file 2: Data S2). This analysis identified two FDA-approved antineoplastic drugs, including everolimus and trametinib, that significantly decreased the expression level of *ERN2*, *ESR1*, *FLT1*, *NR1H4*, *SHOX*, *SP110*, and *ZNF365* (Fig. 3n). Collectively, our analyses demonstrate the MAAS's effectiveness in detecting clinically relevant subpopulations with low CNV heterogeneity but significantly differential therapeutic vulnerabilities.

Moreover, we applied MAAS to a 10x multiome dataset of B-cell lymphoma [23] characterized by minimal CNV burden [106]. This dataset contains paired scRNA-seq and scATAC-seq data measured for each cell. Comparative evaluation against conventional single-modality methods, such as inferCNV [107] and CopyKAT [6] that are widely used to identify tumor cell subpopulations using gene expression-derived copy numbers, as well as dual-modality combination of gene expression and chromatin accessibility (Additional file 1: Fig. S26), demonstrated MAAS's superior capability in partitioning 2,077 malignant cells into nine molecularly distinct clusters (Additional file 1: Fig. S27), despite the limited CNV heterogeneity of this malignancy. Notably, although these clusters shared comparable transcriptional profiles, they displayed marked divergence in SNV and chromatin accessibility profiles (Additional file 1: Fig. S27b-e). Three lines of evidence substantiate their biological distinctness. First, genetic exclusive analysis revealed 181 cluster-specific SNVs distributed across the nine clusters (Additional file 1: Fig. S27b; chi-square test, FDR < 0.05), supporting their classification as genetically distinct subpopulations. Second, epigenetic precedence was evidenced by elevated chromatin accessibility of 3,077 transcriptionally stable genes in cluster 3

(Additional file 1: Fig. S27f-g), in line with established principles of epigenetic regulation mechanisms where chromatin accessibility changes often precedes gene expression changes [108]. Third, the clusters showed significantly functional differences, as we further calculated the E-distance between MAAS clusters with wild-type versus SUMO-activating enzyme inhibitor-treated B-cell lymphoma cell lines, separately (Additional file 1: Supplementary Methods). Cluster 3 exhibited the significantly longest E-distance to the drug-treated subpopulation (Additional file 1: Fig. S27h; Kruskal-Wallis test, *P*-value < 0.01), suggesting intrinsic drug resistance. These findings highlight the value of multimodal analysis in identifying functionally relevant subpopulations that are not apparent from gene expression data alone, particularly in cancer types with subtle CNV differences. Collectively, our analyses demonstrate the MAAS' s effectiveness in detecting clinically relevant subpopulations with low CNV heterogeneity but significantly differential therapeutic vulnerabilities.

MAAS enabled high-resolution identification of a temozolomide-resistant glioma subpopulation

GBM is the most common and aggressive primary brain malignancy in adults [109]. To investigate the heterogeneity of GBM, we applied MAAS to a scATAC-seq dataset of adult GBM [11] (patients CGY4218, CGY4250, CGY4275 and CGY4349) containing 866 tumor cells (Additional file 1: Figs. S28 and S29). MAAS not only recapitulated tumor clusters identified by traditional single-modality approaches, but also resolved a finer subcluster (cluster 2) within a previously defined major cell population at higher resolution (Fig. 4a and b; Additional file 1: Fig. S30). Function enrichment analysis of cluster 2 showed significant enrichment in apoptosis, angiogenesis, and KRAS signaling pathways (Fig. 4c), indicating the

functional heterogeneity within the major cell population. To trace tumor progression, we constructed a minimum-evolution tree that mapped the developmental trajectory of GBM tumor cells (Fig. 4d and e; Additional file 2: Data S3). Notably, cluster 2 exhibited 4,188 DACRs and one cluster-specific SNV (Fig. 4e; Additional file 1: Fig. S31), indicating that chromatin remodeling and mutational events collectively define this unique cluster. Importantly, this SNV was localized within one of the identified DACRs (chr4: 147103187-147104189) (Fig. 4f). Detailed inspection of the chromatin accessibility coverage at this locus further confirmed the co-occurrence of the SNV and DACR (Fig. 4g), supporting a possible interplay between genetic and epigenetic alterations that may contribute to the functional heterogeneity of glioblastoma. These findings were further validated using an independent pediatric GBM (pGBM) dataset [11] (Additional file 1: Fig. S32 and Supplementary Methods).

To evaluate the therapeutic significance of MAAS-identified clusters, we examined their association with temozolomide (TMZ) resistance, the first-line chemotherapeutic for glioma. We first linked the gene activities of each cluster to the half-maximal inhibitory concentration (IC_{50}) of TMZ (Methods). Cluster 2 exhibited significantly greater resistance to TMZ (Fisher's exact test, *P*-value = 2.28×10^{-7}), with accurate predictions supported by the area under the curve (AUC) values of 0.791, 0.733, and 0.694 (Fig. 4h; Additional file 1: Fig. S33a and b). To further validate these findings, we calculated the energy distance (E-distance) [110] between MAAS clusters and TMZ-associated subpopulations from six glioma cell lines [67, 111] (Methods). Cluster 2 consistently showed the shortest E-distance to TMZ-resistant subpopulations (Fig. 4i and Additional file 1: Fig. S33c; Wilcoxon rank-sum test, *P*-value < 0.0001). Additionally, the cluster 2 was further validated using an independent GBM dataset (Additional file 1: Figs. S34 and S35). To experimentally confirm the TMZ resistance of

cluster 2, we conducted knockdown experiments in the U-87 glioblastoma cell line. Downregulation of cluster 2-specific genes, such as *TPST1*, *ADAMTS1*, *RFTN1*, significantly reduced TMZ resistance (Fig. 4j-k; Additional file 1: Fig. S36). Moreover, re-expression of *TPST1* restored TMZ resistance in the rescue experiment (Fig. 4k). Additionally, overexpression experiments further confirmed the role of cluster 2 in conferring TMZ resistance in the U251 cell line (Fig. 4m). In summary, MAAS identified a glioma cell subpopulation with strong TMZ resistance at high resolution, underscoring the potential of the MAAS method as a powerful tool for accurately classifying clinically relevant glioma cell subpopulations.

A new MAAS-derived clinical signature across multiple cancer types

To evaluate the clinical utility of the MAAS method, we developed a new MAAS-derived clinical signature named MAASig, based on the expression of genes identified from subpopulation-specific GRNs (Fig. 5a; Additional file 1: Supplementary Methods). MAASig was constructed by first identifying subpopulation-specific open chromatin accessible regions and marker genes, followed by inferring cis-regulatory links [112] between these regions and their target genes. Transcription factors (TFs) significantly enriched in the DACRs and regulatory links connecting to subpopulation marker genes were retained for GRN construction. Finally, TFs and their target genes from the GRNs of all subpopulations were aggregated to define the candidate signature genes. To develop the prognostic model, we employed an ensemble learning approach with 10-fold cross-validation to prevent overfitting. The most robust features were selected using four complementary feature selection algorithms: LASSO, stepwise Cox regression, CoxBoost and random survival forest. Each cancer-specific model was

validated on independent external cohorts that were not used during model training. In addition to glioblastoma, ovarian cancer and B-cell lymphoma, MAASig was applied to HCC and ccRCC (Additional file 1: Tables S4-S6). Across all five cancer types, MAASig demonstrated significant prognostic value and superior prediction accuracy (Fig. 5b and c; log-rank test, *P*-values = 1.61×10^{-165}, 1.52×10^{-106}, 5.34×10^{-221}, 1.53×10^{-42} and 1.73×10^{-65}, respectively), outperforming traditional clinicopathologic variables and other existing signatures (Fig. 5c; Additional file 1: Figs. S37-S39, Table S7 and Supplementary Methods). For example, in GBM, MAASig achieved an average concordance index (C-index) of 0.861 and a time-dependent AUC of 0.925, significantly outperforming clinical characteristics such as age, *IDH* mutation status, 1p19q copy numbers, and *MGMT* promoter methylation status (Fig. 5b and Additional file 1: Fig. S37). Similarly, in ccRCC, MAASig consistently achieved the highest prediction accuracy, with an average C-index of 0.902 and a time-dependent AUC of 0.925 (Fig. 5b and Additional file 1: Fig. S37). Notably, MAASig remained significantly independent of other clinical features in both training and test sets across the cancer types (Additional file 1: Fig. S39), demonstrating its superiority and robustness in prognosticating patient survival. Calibration plots further confirmed that MAASig was well-calibrated across 1-, 3-, and 5-year time horizons in all cancer types (Additional file 1: Fig. S40). Decision curve analyses consistently demonstrated that MAASig provided a higher net clinical benefit than ‘treating’ everyone or no one across a wide range of threshold probabilities [113] (Additional file 1: Fig. S41).

To further examine the clinical significance, we focused on ccRCC [114], which exhibits substantial intra-tumor heterogeneity that contributes to the drug-tolerance [115]. MAAS accurately identified seven distinct cell subpopulations that were previously overlooked by

traditional methods (Fig. 5d-f and Additional file 1: Fig. S42). To explore potential molecular mechanisms, we examined transcription factors binding motifs in DACRs and inferred TF binding motif activity by estimating the gain or loss of chromatin accessibility (Additional file 1: Supplementary Methods). Our analysis revealed significant variability in TF activity across clusters (Fig. 5g). For example, *NR2C1* activity was elevated in cluster 7, while *CTCF* activity increased in cluster 6. Additionally, we assessed the correlation between the cluster-specific gene modules, identified by weighted correlation network analysis (WGCNA) [116], and immunotherapeutic sensitivity (Additional file 1: Supplementary Methods). We found that cluster 7, represented by gene module 4, demonstrated the highest resistance to anti-PD-1 blockade therapy with nivolumab (Fig. 5h-j; Additional file 1: Figs. S43 and S44). This finding was further validated using multiple immunotherapy response metrics, including tumor immune dysfunction and exclusion (TIDE) score [117], MHC-I association immunoscore (MIAS) [118], 18-gene gene expression profile (GEP) [119], and PD-1 gene activity [120] (Fig. 5k-n). In summary, the MAAS-derived signature shows strong prognostic value and robustness in predicting patient survival across multiple cancer types.

Discussion

To our knowledge, MAAS is the first computational method for multimodal integration of scATAC-seq data capable of identifying critical tumor cell subpopulations distinct from those determined by traditional single-modality approaches, such as Copy-scAT [11] and epiAneufinder [12]. MAAS addresses several key limitations of existing approaches. First, it incorporates rigorous normalization and data denoising procedures to mitigate potential technical confounders such as variable cell coverage. Our extensive benchmarking demonstrates that the subpopulations

identified by MAAS represent genuine biological differences rather than technical artifacts. Second, we found that MAAS provides higher accuracy in identifying tumor subpopulations compared to other available methods. By integrating multimodal data, MAAS uncovers new tumor cell subpopulations with significant biological and clinical relevance. The MAAS method is fundamentally different from previous subpopulation prediction methods. Instead of relying solely on single-modality features, which often overlook crucial layers of epigenomic information, MAAS maximizes the extraction of informative features from scATAC-seq data. Additionally, its self-expressive multimodal matrix factorization strategy enhances multimodal signals, enabling a more robust classification of tumor subpopulations. Furthermore, MAAS is an explainable multimodal integration method that quantifies the contribution of each modality to cell cluster assignment. For example, the two pediatric ependymoma cell subpopulations predicted by MAAS were primarily driven by chromatin accessibility and SNVs, which contributed 69.68% and 66.49% more than CNVs, respectively (Fig. 3c). The feasibility of the MAAS method is also noteworthy, as it allows for the simultaneous examination of genetic mutations and epigenetic variations without requiring additional single-cell assays. Despite the improved accuracy of MAAS, it requires increased computational time. This limitation, however, may be alleviated through the implementation of distributed or heuristic algorithms. Moreover, since MAAS was specifically developed for tumor cells, it may not be suitable for normal or non-malignant cell populations, where the underlying biological assumptions and patterns of heterogeneity may differ substantially.

Importantly, beyond its methodological advances, MAAS provides novel biological insights into tumor heterogeneity. For example, MAAS resolved a high-resolution glioma subpopulation with strong TMZ resistance, a major clinical obstacle that contributes to

therapeutic failure and tumor recurrence. The ability of MAAS to resolve such clinically relevant subpopulations demonstrates that multimodal integration can go beyond cell classification to uncover functional heterogeneity with direct therapeutic implications. Moreover, genes highly expressed in this cluster, especially *TPST1*, was experimentally validated with TMZ resistance. Previous studies demonstrated that *TPST1* can mediate the tyrosine sulfation of the chemokine receptor *CXCR4*, thereby enhancing *CXCL12*/*CXCR4*-dependent signaling and promoting tumor cell migration and invasiveness [121]. Moreover, recent work on *TPST2* revealed that tyrosine sulfation can modulate immune-related receptors and affect the tumor cell response to interferon signaling and anti-PD1 treatment [122]. By analogy, *TPST1* activity may also shape the tumor-immune microenvironment and influence therapeutic outcomes beyond chemotherapy. These findings highlight the potential of *TPST1* as novel therapeutic targets, and suggest that pharmacological inhibition of *TPST1* or interference with sulfation-dependent signaling could help overcome TMZ resistance. Additionally, in pediatric ependymoma, where CNV burdens are extremely low, MAAS successfully resolved a highly proliferative, chemoresistant subpopulation and nominated everolimus and trametinib as potential therapeutic agents. While a Phase II study reported that everolimus did not show significant anti-tumor activity in this context [123], trametinib has demonstrated clinical activity in pediatric CNS tumors, achieving disease control or partial responses in a subset of patients with MAPK pathway activation [124]. These findings suggest that targeting the aggressive subpopulation identified by MAAS with trametinib may offer a promising therapeutic strategy. Notably, the cell-line models used in our study define TMZ resistance primarily based on IC_{50} values. This threshold has been well established in the literature as

a proxy for resistance [67], but we acknowledge that clinical resistance is a more complex phenomenon that involves a variety of factors, such as aberrant signaling pathways, autophagy, epigenetic modifications, and extracellular vesicle production [125]. In clinical settings, TMZ resistance is often defined by factors such as tumor recurrence during treatment or progression-free survival time [126]. Therefore, while the cell-line resistance models provide valuable insights into potential mechanisms of resistance, we recommend that future studies validate these findings in clinical patient samples to further confirm the relevance of these *in vitro* models to clinical resistance.

Motivated by the biological and therapeutic insights uncovered by MAAS, we further investigated its potential clinical utility in predicting patient outcomes. Prognostic markers are clinical measures that predict patient outcomes, such as recurrence or survival, and range from simple anatomical features to complex molecular indicators reflecting underlying disease biology [127]. However, most existing prognostic signatures have been derived from bulk-level data, which obscure the profound intratumoral heterogeneity that exists among distinct tumor subpopulations differing in metabolic activity, survival signaling, and epigenetic regulation [2]. To overcome this limitation, we integrated subpopulation-specific molecular features identified by MAAS to construct a multimodal prognostic signature, MAASig. By explicitly incorporating information from functionally distinct tumor subpopulations, MAASig captures biologically relevant heterogeneity and exhibits significantly improved predictive performance compared with conventional clinicopathological features and previously reported signatures. Beyond its methodological importance, MAASig also enables more precise patient stratification, supports risk-adapted treatment planning, and may guide the selection of targeted therapies for patients most likely to benefit. Furthermore, MAASig could be integrated with existing clinical workflows

as a complementary biomarker to enhance prognostic accuracy and inform personalized therapeutic decision-making.

Conclusions

In summary, the MAAS method underscores the power of multimodal integration in dissecting tumor heterogeneity using single-cell epigenomics data. MAAS will enable the broad application of widely available single-cell sequencing data in oncology and other diseases, ultimately revealing critical cell subpopulations for cell-targeted treatments.

References

1. Mazor T, Pankov A, Johnson BE, Hong C, Hamilton EG, Bell RJA, Smirnov IV, Reis GF, Phillips JJ, Barnes MJ, et al: **DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors.** *Cancer Cell* 2015, **28:**307-317.
2. Dagogo-Jack I, Shaw AT: **Tumour heterogeneity and resistance to cancer therapies.** *Nat Rev Clin Oncol* 2018, **15:**81-94.
3. Sun D, Guan X, Moran AE, Wu LY, Qian DZ, Schedin P, Dai MS, Danilov AV, Alumkal JJ, Adey AC, et al: **Identifying phenotype-associated subpopulations by integrating bulk and single-cell sequencing data.** *Nat Biotechnol* 2022, **40:**527-538.
4. Zhao J, Jaffe A, Li H, Lindenbaum O, Sefik E, Jackson R, Cheng X, Flavell RA, Kluger Y: **Detection of differentially abundant cell subpopulations in scRNA-seq data.** *Proc Natl Acad Sci U S A* 2021, **118**.
5. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir el AD, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER, et al: **Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis.** *Cell* 2015, **162:**184-197.
6. Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, Kumar T, Hu M, Sei E, Davis A, et al: **Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes.** *Nat Biotechnol* 2021, **39:**599-608.
7. Zhou Z, Xu B, Minn A, Zhang NR: **DENDRO: genetic heterogeneity profiling and subclone detection by single-cell RNA sequencing.** *Genome Biol* 2020, **21:**10.

8. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard JK, Kundaje A, Greenleaf WJ, et al: **Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution.** *Nat Genet* 2016, **48:**1193-1203.
9. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, Silva TC, Groeneveld C, Wong CK, Cho SW, et al: **The chromatin accessibility landscape of primary human cancers.** *Science* 2018, **362**.
10. Buenrostro JD, Corces MR, Lareau CA, Wu B, Schep AN, Aryee MJ, Majeti R, Chang HY, Greenleaf WJ: **Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation.** *Cell* 2018, **173:**1535-1548 e1516.
11. Nikolic A, Singhal D, Ellestad K, Johnston M, Shen Y, Gillmor A, Morrissy S, Cairncross JG, Jones S, Lupien M, et al: **Copy-scAT: Deconvoluting single-cell chromatin accessibility of genetic subclones in cancer.** *Sci Adv* 2021, **7:**eabg6045.
12. Ramakrishnan A, Symeonidi A, Hanel P, Schmid KT, Richter ML, Schubert M, Colome-Tatche M: **epiAneufinder identifies copy number alterations from single-cell ATAC-seq data.** *Nat Commun* 2023, **14:**5846.
13. Wu CY, Lau BT, Kim HS, Sathe A, Grimes SM, Ji HP, Zhang NR: **Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer.** *Nat Biotechnol* 2021, **39:**1259-1269.
14. Choi JD, Lee JS: **Interplay between Epigenetics and Genetics in Cancer.** *Genomics Inform* 2013, **11:**164-173.
15. Nam AS, Chaligne R, Landau DA: **Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics.** *Nat Rev Genet* 2021, **22:**3-18.
16. De Falco A, Caruso F, Su XD, Iavarone A, Ceccarelli M: **A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data.** *Nat Commun* 2023, **14:**1074.
17. Chen H, Lareau C, Andreani T, Vinyard ME, Garcia SP, Clement K, Andrade-Navarro MA, Buenrostro JD, Pinello L: **Assessment of computational methods for the analysis of single-cell ATAC-seq data.** *Genome Biol* 2019, **20:**241.
18. Queitsch K, Moore TW, O'Connell BL, Nichols RV, Muschler JL, Keith D, Lopez C, Sears RC, Mills GB, Yardimci GG, Adey AC: **Accessible high-throughput single-cell whole-genome sequencing with paired chromatin accessibility.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE243430; 2023.
19. Wu CY, Lau BT, Kim HS, Sathe A, Grimes SM, Ji HP, Zhang NR: **Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer.** Dataset. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA674903/ and https://www.ncbi.nlm.nih.gov/bioproject/PRJNA498809/; 2021.
20. Regner MJ, Wisniewska K, Garcia-Recio S, Thennavan A, Mendez-Giraldez R, Malladi VS, Hawkins G, Parker JS, Perou CM, Bae-Jump VL, Franco HL: **A multi-omic single-cell landscape of human**

gynecologic malignancies. Dataset. database of Genotypes and Phenotypes. https://dbgap.ncbi.nlm.nih.gov/beta/search/?OBJ=study&TERM=phs002340.v1.p1; 2021.

21. Croft W, Pounds R, Jeevan D, Singh K, Balega J, Sundar S, Williams A, Ganesan R, Kehoe S, Ott S, et al: **The chromatin landscape of high-grade serous ovarian cancer metastasis identifies regulatory drivers in post-chemotherapy residual tumour cells.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE247982; 2024.
22. Aubin RG, Troisi EC, Montelongo J, Alghalith AN, Nasrallah MP, Santi M, Camara PG: **Pro-inflammatory cytokines mediate the epithelial-to-mesenchymal-like transition of pediatric posterior fossa ependymoma.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE206579; 2022.
23. **Flash-Frozen Lymph Node with B Cell Lymphoma (14k sorted nuclei) - Epi Multiome ATAC + Gene Expression dataset analyzed using Cell Ranger ARC 2.0.0.** *10X Genomics https://www10xgenomicscom/datasets/fresh-frozen-lymph-node-with-b-cell-lymphoma-14-k-sorted-nuclei-1-standard-2-0-0* 2021.
24. Nikolic A, Singhal D, Ellestad K, Johnston M, Shen Y, Gillmor A, Morrissy S, Cairncross JG, Jones S, Lupien M, et al: **Copy-scAT: Deconvoluting single-cell chromatin accessibility of genetic subclones in cancer.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE163655 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE139136; 2021.
25. Craig AJ, Silveira MAD, Ma L, Revsine M, Wang L, Heinrich S, Rae Z, Ruchinskas A, Dadkhah K, Do W, et al: **Genome-wide profiling of transcription factor activity in primary liver cancer using single-cell ATAC sequencing.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE227265; 2023.
26. Yu Z, Lv Y, Su C, Lu W, Zhang R, Li J, Guo B, Yan H, Liu D, Yang Z, et al: **Integrative Single-Cell Analysis Reveals Transcriptional and Epigenetic Regulatory Features of Clear Cell Renal Cell Carcinoma.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE207493; 2023.
27. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, Olsen BN, Mumbach MR, Pierce SE, Corces MR, et al: **Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion.** Dataset. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA533341/; 2019.
28. Henriquez NV, Forshew T, Tatevossian R, Ellis M, Richard-Loendt A, Rogers H, Jacques TS, Reitboeck PG, Pearce K, Sheer D, et al: **Comparative expression analysis reveals lineage relationships between human and murine gliomas and a dominance of glial signatures during tumor propagation in vitro.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42658; 2013.

29. Van MT, Broaddus W, Dumur C: **Frozen tumor ependymoma biopsy samples from pediatric patients.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13267; 2018.

30. Griesinger AM, Josephson RJ, Donson AM, Mulcahy Levy JM, Amani V, Birks DK, Hoffman LM, Furtek SL, Reigan P, Handler MH, et al: **Interleukin-6/STAT3 Pathway Signaling Drives an Inflammatory Phenotype in Group A Ependymoma.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66354; 2015.

31. Hiddingh L, Tannous BA, Teng J, Tops B, Jeuken J, Hulleman E, Boots-Sprenger SH, Vandertop WP, Noske DP, Kaspers GJ, et al: **EFEMP1 induces gamma-secretase/Notch-mediated temozolomide resistance in glioblastoma.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53014; 2014.

32. Tso JL, Yang S, Menjivar JC, Yamada K, Zhang Y, Hong I, Bui Y, Stream A, McBride WH, Liau LM, et al: **Bone morphogenetic protein 7 sensitizes O6-methylguanine methyltransferase expressing-glioblastoma stem cells to clinically relevant dose of temozolomide.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68029; 2015.

33. Zhao Z, Zhang KN, Wang Q, Li G, Zeng F, Zhang Y, Wu F, Chai R, Wang Z, Zhang C, et al: **Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data from Chinese Glioma Patients.** *Genomics Proteomics Bioinformatics* 2021, **19:**1-12.

34. Lam V, Bruss N, Liu T, Danilov AV: **Pharmacologic targeting of SUMOylation drives mitochondrial dysfunction and metabolic alterations in B cell Malignancies.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE254913; 2024.

35. Gao GF, Parker JS, Reynolds SM, Silva TC, Wang LB, Zhou W, Akbani R, Bailey M, Balu S, Berman BP, et al: **Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data.** *Cell Syst* 2019, **9:**24-34 e10.

36. Kommoss S, Winterhoff B, Oberg AL, Konecny GE, Wang C, Riska SM, Fan JB, Maurer MJ, April C, Shridhar V, et al: **Bevacizumab May Differentially Improve Ovarian Cancer Outcome in Patients with Proliferative and Mesenchymal Molecular Subtypes.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140082; 2017.

37. Yoshihara K, Tsunoda T, Shigemizu D, Fujiwara H, Hatae M, Fujiwara H, Masuzaki H, Katabuchi H, Kawakami Y, Okamoto A, et al: **High-risk ovarian cancer based on 126-gene expression signature is uniquely characterized by downregulation of antigen presentation pathway.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32062; 2012.

38. Painter D, Barrans S, Lacy S, Smith A, Crouch S, Westhead D, Sha C, Patmore R, Tooze R, Burton C, Roman E: **Cell-of-origin in diffuse large B-cell lymphoma: findings from the UK's population-based Haematological Malignancy Research Network.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE181063; 2019.

39. Lenz G, Wright G, Dave SS, Xiao W, Powell J, Zhao H, Xu W, Tan B, Goldschmidt N, Iqbal J, et al: **Stromal gene signatures in large-B-cell lymphomas.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10846; 2008.
40. Dubois S, Tesson B, Mareschal S, Viailly PJ, Bohers E, Ruminy P, Etancelin P, Peyrouze P, Copie-Bergman C, Fabiani B, et al: **Refining diffuse large B-cell lymphoma subgroups using integrated analysis of molecular profiles.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE136971; 2019.
41. Ning J, Ye Y, Shen H, Zhang R, Li H, Song T, Zhang R, Liu P, Chen G, Wang H, et al: **Macrophage-coated tumor cluster aggravates hepatoma invasion and immunotherapy resistance via generating local immune deprivation.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116174; 2024.
42. Grinchuk OV, Yenamandra SP, Iyer R, Singh M, Lee HK, Lim KH, Chow PK, Kuznetsov VA: **Tumor-adjacent tissue co-expression profile analysis reveals pro-oncogenic ribosomal gene signature for prognosis of resectable hepatocellular carcinoma.** Dataset. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76427; 2018.
43. Sato Y, Yoshizato T, Shiraishi Y, Maekawa S, Okuno Y, Kamura T, Shimamura T, Sato-Otsubo A, Nagae G, Suzuki H, et al: **Integrated molecular analysis of clear-cell renal cell carcinoma.** Dataset. European Bioinformatics Institute. https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-1980; 2013.
44. Clark DJ, Dhanasekaran SM, Petralia F, Pan J, Song X, Hu Y, da Veiga Leprevost F, Reva B, Lih TM, Chang HY, et al: **Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma.** Dataset. LinkedOmics. https://kb.linkedomics.org/download; 2019.
45. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C: **The sequence read archive.** *Nucleic Acids Res* 2011, **39:**D19-21.
46. **cell-ranger-atac** [https://www.10xgenomics.com/support/cn/software/cell-ranger-atac/2.1/release-notes/installation]
47. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R: **Single-cell chromatin state analysis with Signac.** *Nat Methods* 2021, **18:**1333-1341.
48. Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, DeWitt WS, et al: **A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility.** *Cell* 2018, **174:**1309-1324 e1318.
49. Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods.** *Biostatistics* 2007, **8:**118-127.
50. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD: **The sva package for removing batch effects and other unwanted variation in high-throughput experiments.** *Bioinformatics* 2012, **28:**882-883.
51. Muyas F, Sauer CM, Valle-Inclan JE, Li R, Rahbari R, Mitchell TJ, Hormoz S, Cortes-Ciriano I: **De novo detection of somatic mutations in high-throughput single-cell profiling data sets.** *Nat Biotechnol* 2024, **42:**758-767.

52. Yan J, Xi J, Yu Z: **A parametric model for clustering single-cell mutation data.** In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. pp. 253-260; 2022:253-260.
53. Rousseeuw PJ: **Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.** *Journal of Computational and Applied Mathematics* 1987, **20:**53-65.
54. Davies DL, Bouldin DW: **A Cluster Separation Measure.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1979, **PAMI-1:**224-227.
55. Pakhira MK, Bandyopadhyay S, Maulik U: **Validity index for crisp and fuzzy clusters.** *Pattern Recognition* 2004, **37:**487-501.
56. Caliński T, Harabasz J: **A dendrite method for cluster analysis.** *Communications in Statistics* 1974, **3:**1-27.
57. Rzhetsky A, Nei M: **Theoretical foundation of the minimum-evolution method of phylogenetic inference.** *Mol Biol Evol* 1993, **10:**1073-1095.
58. Paradis E, Claude J, Strimmer K: **APE: Analyses of Phylogenetics and Evolution in R language.** *Bioinformatics* 2004, **20:**289-290.
59. Yu G: **Using ggtree to Visualize Data on Tree-Like Structures.** *Curr Protoc Bioinformatics* 2020, **69:**e96.
60. Li C, Shao X, Zhang S, Wang Y, Jin K, Yang P, Lu X, Fan X, Wang Y: **scRank infers drug-responsive cell types from untreated scRNA-seq data using a target-perturbed gene regulatory network.** *Cell Rep Med* 2024, **5:**101568.
61. Yerly L, Andreatta M, Garnica J, Domizio JD, Gilliet M, Carmona SJ, Kuonen F: **Wounding triggers invasive progression in human basal cell carcinoma.** *bioRxiv* 2024**:**2024.2005.2031.596823.
62. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R: **Integrating single-cell transcriptomic data across different conditions, technologies, and species.** *Nat Biotechnol* 2018, **36:**411-420.
63. Maeser D, Gruener RF, Huang RS: **oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data.** *Brief Bioinform* 2021, **22**.
64. Griesinger AM, Josephson RJ, Donson AM, Mulcahy Levy JM, Amani V, Birks DK, Hoffman LM, Furtek SL, Reigan P, Handler MH, et al: **Interleukin-6/STAT3 Pathway Signaling Drives an Inflammatory Phenotype in Group A Ependymoma.** *Cancer Immunol Res* 2015, **3:**1165-1174.
65. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, et al: **Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells.** *Nucleic Acids Res* 2013, **41:**D955-961.
66. **Rfast: A Collection of Efficient and Extremely Fast R Functions** [https://CRAN.R-project.org/package=Rfast]
67. Hiddingh L, Tannous BA, Teng J, Tops B, Jeuken J, Hulleman E, Boots-Sprenger SH, Vandertop WP, Noske DP, Kaspers GJ, et al: **EFEMP1 induces gamma-secretase/Notch-mediated temozolomide resistance in glioblastoma.** *Oncotarget* 2014, **5:**363-374.

68. Clough E, Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, et al: **NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update.** *Nucleic Acids Res* 2024, **52:**D138-D144.
69. Chen H, Wang Z, Gong L, Wang Q, Chen W, Wang J, Ma X, Ding R, Li X, Zou X, et al: **A distinct class of pan-cancer susceptibility genes revealed by an alternative polyadenylation transcriptome-wide association study.** *Nat Commun* 2024, **15:**1729.
70. Zou Y, Liu Z, Zhou Y, Wang J, Xu Q, Zhao X, Miao Z: **TRPC5 mediates TMZ resistance in TMZ-resistant glioblastoma cells via NFATc3-P-gp pathway.** *Transl Oncol* 2021, **14:**101214.
71. Mann HB, Whitney DR: **On a test of whether one of two random variables is stochastically larger than the other.** *The annals of mathematical statistics* 1947**:**50-60.
72. Pearson K: **X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material.** *Philosophical Transactions of the Royal Society of London(A)* 1895**:**343-414.
73. Fisher RA: **On the interpretation of χ 2 from contingency tables, and the calculation of P.** *Journal of the royal statistical society* 1922, **85:**87-94.
74. Kaplan EL, Meier P: **Nonparametric estimation from incomplete observations.** *Journal of the American statistical association* 1958, **53:**457-481.
75. Therneau T.M. GPM: *Modeling Survival Data: Extending the Cox Model.* Springer, New York; 2000.
76. Ward MD, Ahlquist JS: *Maximum likelihood for social science: Strategies for analysis.* Cambridge University Press; 2018.
77. **survivalROC: Time-Dependent ROC Curve Estimation from Censored Survival Data** [https://CRAN.R-project.org/package=survivalROC]
78. Wu CY, Lau BT, Kim HS, Sathe A, Grimes SM, Ji HP, Zhang NR: **Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer.** *Nat Biotechnol* 2021, **39:**1259-1269.
79. Zhang J, Li C, You C, Qi X, Zhang H, Guo J, Lin Z: **Self-Supervised Convolutional Subspace Clustering Network.** In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; *15-20 June 2019*. 2019: 5468-5477.
80. McInnes L, Healy J, Melville J: **Umap: Uniform manifold approximation and projection for dimension reduction.** *arXiv preprint arXiv:180203426* 2018.
81. Chalise P, Fridley BL: **Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm.** *PLoS One* 2017, **12:**e0176278.
82. Pierre-Jean M, Mauger F, Deleuze JF, Le Floch E: **PIntMF: Penalized Integrative Matrix Factorization method for multi-omics data.** *Bioinformatics* 2022, **38:**900-907.
83. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A: **Similarity network fusion for aggregating data types on a genomic scale.** *Nat Methods* 2014, **11:**333-337.

84. Wu D, Wang D, Zhang MQ, Gu J: **Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification.** *BMC Genomics* 2015, **16:**1022.

85. Meng C, Kuster B, Culhane AC, Gholami AM: **A multivariate approach to the integration of multi-omics datasets.** *BMC Bioinformatics* 2014, **15:**162.

86. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, Stegle O: **MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data.** *Genome Biol* 2020, **21:**111.

87. Ashuach T, Gabitto MI, Koodli RV, Saldi GA, Jordan MI, Yosef N: **MultiVI: deep generative model for the integration of multimodal data.** *Nat Methods* 2023, **20:**1222-1231.

88. Cheng M, Li Z, Costa IG: **MOJITOO: a fast and universal method for integration of multimodal single-cell data.** *Bioinformatics* 2022, **38:**i282-i289.

89. Persad S, Choo ZN, Dien C, Sohail N, Masilionis I, Chaligne R, Nawy T, Brown CC, Sharma R, Pe'er I, et al: **SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data.** *Nat Biotechnol* 2023, **41:**1746-1757.

90. Li Z, Kuppe C, Ziegler S, Cheng M, Kabgani N, Menzel S, Zenke M, Kramann R, Costa IG: **Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen.** *Nat Commun* 2021, **12:**6386.

91. Johnson JAI, Tsang AP, Mitchell JT, Zhou DL, Bowden J, Davis-Marcisak E, Sherman T, Liefeld T, Loth M, Goff LA, et al: **Inferring cellular and molecular processes in single-cell data with non-negative matrix factorization using Python, R and GenePattern Notebook implementations of CoGAPS.** *Nat Protoc* 2023, **18:**3690-3731.

92. Smith MR: **Information theoretic generalized Robinson-Foulds metrics for comparing phylogenetic trees.** *Bioinformatics* 2020, **36:**5007-5013.

93. Queitsch K, Moore TW, O'Connell BL, Nichols RV, Muschler JL, Keith D, Lopez C, Sears RC, Mills GB, Yardimci GG, Adey AC: **Accessible high-throughput single-cell whole-genome sequencing with paired chromatin accessibility.** *Cell Rep Methods* 2023, **3:**100625.

94. Regner MJ, Wisniewska K, Garcia-Recio S, Thennavan A, Mendez-Giraldez R, Malladi VS, Hawkins G, Parker JS, Perou CM, Bae-Jump VL, Franco HL: **A multi-omic single-cell landscape of human gynecologic malignancies.** *Mol Cell* 2021, **81:**4924-4941 e4910.

95. Croft W, Pounds R, Jeevan D, Singh K, Balega J, Sundar S, Williams A, Ganesan R, Kehoe S, Ott S, et al: **The chromatin landscape of high-grade serous ovarian cancer metastasis identifies regulatory drivers in post-chemotherapy residual tumour cells.** *Commun Biol* 2024, **7:**1211.

96. Yang Y, Yang L: **Somatic structural variation signatures in pediatric brain tumors.** *Cell Rep* 2023, **42:**113276.

97. Aubin RG, Troisi EC, Montelongo J, Alghalith AN, Nasrallah MP, Santi M, Camara PG: **Pro-inflammatory cytokines mediate the epithelial-to-mesenchymal-like transition of pediatric posterior fossa ependymoma.** *Nat Commun* 2022, **13:**3936.

98. Rojo de la Vega M, Chapman E, Zhang DD: **NRF2 and the Hallmarks of Cancer.** *Cancer Cell* 2018, **34:**21-43.

99. Jurikova M, Danihel L, Polak S, Varga I: **Ki67, PCNA, and MCM proteins: Markers of proliferation in the diagnosis of breast cancer.** *Acta Histochem* 2016, **118:**544-552.

100. Aiello NM, Kang Y: **Context-dependent EMT programs in cancer metastasis.** *J Exp Med* 2019, **216:**1016-1026.

101. Katoh M: **Function and cancer genomics of FAT family genes (review).** *Int J Oncol* 2012, **41:**1913-1918.

102. Henriquez NV, Forshew T, Tatevossian R, Ellis M, Richard-Loendt A, Rogers H, Jacques TS, Reitboeck PG, Pearce K, Sheer D, et al: **Comparative expression analysis reveals lineage relationships between human and murine gliomas and a dominance of glial signatures during tumor propagation in vitro.** *Cancer Res* 2013, **73:**5834-5844.

103. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, Zhang F, Mundlos S, Christiansen L, Steemers FJ, et al: **The single-cell transcriptional landscape of mammalian organogenesis.** *Nature* 2019, **566:**496-502.

104. Li C, Shao X, Zhang S, Wang Y, Jin K, Yang P, Lu X, Fan X, Wang Y: **scRank infers drug-responsive cell types from untreated scRNA-seq data using a target-perturbed gene regulatory network.** *Cell Rep Med* 2024**:**101568.

105. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, et al: **A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles.** *Cell* 2017, **171:**1437-1452 e1417.

106. Steele CD, Abbasi A, Islam SMA, Bowes AL, Khandekar A, Haase K, Hames-Fathi S, Ajayi D, Verfaillie A, Dhami P, et al: **Signatures of copy number alterations in human cancer.** *Nature* 2022, **606:**984-991.

107. Liau BB, Sievers C, Donohue LK, Gillespie SM, Flavahan WA, Miller TE, Venteicher AS, Hebert CH, Carey CD, Rodig SJ, et al: **Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance.** *Cell Stem Cell* 2017, **20:**233-246 e237.

108. Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, Ding J, Brack A, Kartha VK, Tay T, et al: **Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin.** *Cell* 2020, **183:**1103-1116 e1120.

109. Hara T, Chanoch-Myers R, Mathewson ND, Myskiw C, Atta L, Bussema L, Eichhorn SW, Greenwald AC, Kinker GS, Rodman C, et al: **Interactions between cancer cells and immune cells drive transitions to mesenchymal-like states in glioblastoma.** *Cancer Cell* 2021, **39:**779-792 e711.

110. Székely GJ, Rizzo ML: **Energy statistics: A class of statistics based on distances.** *Journal of Statistical Planning and Inference* 2013, **143:**1249-1272.

111. Tso JL, Yang S, Menjivar JC, Yamada K, Zhang Y, Hong I, Bui Y, Stream A, McBride WH, Liau LM, et al: **Bone morphogenetic protein 7 sensitizes O6-methylguanine methyltransferase expressing-glioblastoma stem cells to clinically relevant dose of temozolomide.** *Mol Cancer* 2015, **14:**189.

112. Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, Srivatsan S, Qiu X, Jackson D, Minkina A, et al: **Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data.** *Mol Cell* 2018, **71:**858-871 e858.

113. Rousson V, Zumbrunn T: **Decision curve analysis revisited: overall net benefit, relationships to ROC curve analysis, and application to case-control studies.** *BMC Med Inform Decis Mak* 2011, **11:**45.

114. Yu Z, Lv Y, Su C, Lu W, Zhang R, Li J, Guo B, Yan H, Liu D, Yang Z, et al: **Integrative Single-Cell Analysis Reveals Transcriptional and Epigenetic Regulatory Features of Clear Cell Renal Cell Carcinoma.** *Cancer Res* 2023, **83:**700-719.

115. Posadas EM, Limvorasak S, Figlin RA: **Targeted therapies for renal cell carcinoma.** *Nat Rev Nephrol* 2017, **13:**496-511.

116. Morabito S, Reese F, Rahimzadeh N, Miyoshi E, Swarup V: **hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data.** *Cell Rep Methods* 2023, **3:**100498.

117. Jiang P, Gu S, Pan D, Fu J, Sahu A, Hu X, Li Z, Traugh N, Bu X, Li B, et al: **Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response.** *Nat Med* 2018, **24:**1550-1558.

118. Wu CC, Wang YA, Livingston JA, Zhang J, Futreal PA: **Prediction of biomarkers and therapeutic combinations for anti-PD-1 immunotherapy using the global gene network association.** *Nat Commun* 2022, **13:**42.

119. Cristescu R, Mogg R, Ayers M, Albright A, Murphy E, Yearley J, Sher X, Liu XQ, Lu H, Nebozhyn M, et al: **Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy.** *Science* 2018, **362**.

120. Jiang Y, Chen M, Nie H, Yuan Y: **PD-1 and PD-L1 in cancer immunotherapy: clinical implications and future considerations.** *Hum Vaccin Immunother* 2019, **15:**1111-1122.

121. Xu J, Deng X, Tang M, Li L, Xiao L, Yang L, Zhong J, Bode AM, Dong Z, Tao Y, Cao Y: **Tyrosylprotein sulfotransferase-1 and tyrosine sulfation of chemokine receptor 4 are induced by Epstein-Barr virus encoded latent membrane protein 1 and associated with the metastatic potential of human nasopharyngeal carcinoma.** *PLoS One* 2013, **8:**e56114.

122. Oh Y, Kim S, Kim Y, Kim H, Jang D, Shin S, Lee SJ, Kim J, Lee SE, Oh J, et al: **Genome-wide CRISPR screening identifies tyrosylprotein sulfotransferase-2 as a target for augmenting anti-PD1 efficacy.** *Mol Cancer* 2024, **23:**155.

123. Bowers DC, Rajaram V, Karajannis MA, Gardner SL, Su JM, Baxter P, Partap S, Klesse LJ: **Phase II study of everolimus for recurrent or progressive pediatric ependymoma.** *Neurooncol Adv* 2023, **5:**vdad011.

124. Perreault S, Larouche V, Tabori U, Hawkin C, Lippe S, Ellezam B, Decarie JC, Theoret Y, Metras ME, Sultan S, et al: **A phase 2 study of trametinib for patients with pediatric glioma or plexiform neurofibroma with refractory tumor and activation of the MAPK/ERK pathway: TRAM-01.** *BMC Cancer* 2019, **19:**1250.

125. Singh N, Miner A, Hennis L, Mittal S: **Mechanisms of temozolomide resistance in glioblastoma - a comprehensive review.** *Cancer Drug Resist* 2021, **4:**17-43.

126. Ortiz R, Perazzoli G, Cabeza L, Jimenez-Luna C, Luque R, Prados J, Melguizo C: **Temozolomide: An Updated Overview of Resistance Mechanisms, Nanotechnology Advances and Clinical Applications.** *Curr Neuropharmacol* 2021, **19:**513-537.

127. Galon J, Angell HK, Bedognetti D, Marincola FM: **The continuum of cancer immunosurveillance: prognostic, predictive, and mechanistic signatures.** *Immunity* 2013, **39:**11-26.

128. Xiong K, Ding R, Li L: **Multimodal-based analysis of single-cell ATAC-seq data enables highly accurate delineation of clinically relevant tumor cell subpopulations.** GitHub. https://github.com/Larrycpan/MAAS; 2025.

List of abbreviations

scATAC-seq: single-cell assay for transposase-accessible chromatin using sequencing

scRNA-seq: single-cell RNA sequencing

MAAS: Multimodal-based Analysis of scATAC-Seq data

CNV: copy number variation

SNV: single-nucleotide variants

DACR: differentially accessible chromatin region

GRN: gene regulatory network

IC50: half maximal inhibitory concentration

PPFE: pediatric posterior fossa ependymoma

GBM: glioblastoma

TMZ: temozolomide

OC: ovarian cancer

HCC: hepatocellular carcinoma

ccRCC: clear cell renal cell carcinoma

TF-IDF: term frequency-inverse document frequency

UMAP: uniform manifold approximation and projection

WGCNA: weighted correlation network analysis

TF: transcription factor

TIDE: tumor immune dysfunction and exclusion

GEP: gene expression profile

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Funding

This work was funded by grants from the Shenzhen Medical Research Fund (grant no. C2503001 to L.L.), the Major Program of Shenzhen Bay Laboratory (grant no. C1012524001 to L.L.), and the National Natural Science Foundation of China (grant no. 32370721, 32100533 to L.L., grant no. 32570730 to X.Z.).

Availability of data and materials

The raw glioma, PPFE, ccRCC and HCC scATAC-seq data are available in the NCBI database under accession numbers GSE139136 (GBM) [24], GSE163655 (pGBM) [24], GSE206579 [22], GSE207493 [26], and GSE227265 [25], respectively. The raw ovarian cancer scATAC-seq data with pathological stage and metastatic status is available via the database of Genotypes and Phenotypes (dbGaP) under the accession number phs002340.v1.p1 [20]. The raw ovarian cancer scATAC-seq data with treatment information is available in the NCBI database under the accession number GSE247982 [21]. The raw B-cell lymphoma scATAC-seq data are available from 10x Genomics [23]. The scATAC-seq dataset for the SNU601 cell line is available from the NCBI database under the accession number PRJNA674903 [19], and the single-cell whole-genome sequencing data for the SNU601 cell line is available under the accession number PRJNA498809 [19]. The tumor samples of patients SU006 and SU008 are available in the NCBI database under the accession number PRJNA533341 [27]. The single-cell K562 dataset is available under the accession number GSE243430 [18]. The bulk RNA-seq and clinical information of pediatric ependymoma is available in the NCBI under the accession number GSE42658 [28]. Gene expression and clinical features of patients from TCGA cohort are available from the GDC portal (https://portal.gdc.cancer.gov/) [35]. Gene expression profiles of patients from the two cohorts CGGA693 and CGGA325 are publicly available from the Chinese Glioma Genome Atlas (https://www.cgga.org.cn/) [33]. Gene expression of experimentally determined wild-type and TMZ-resistant glioma cells were obtained from the NCBI database under the accession numbers GSE53014 [31] and GSE68029 [32]. Two bulk RNA-seq datasets of PPFE were obtained from the NCBI database under the accession numbers GSE13267 [29] and GSE66354 [30]. Bulk ATAC-seq data of B-cell lymphoma cell lines were obtained from the NCBI database under the accession number GSE254913 [34]. Datasets used for clinical signature

analysis from the NCBI database are available under the following accession numbers: (1) ovarian cancer, GSE140082 [36], and GSE32062 [37]; (2) B-cell lymphoma, GSE181063 [38], GSE10846 [39], and GSE136971 [40]; (3) hepatocellular carcinoma, GSE116174 [41], and GSE76427 [42]. Gene expression and clinical information of ccRCC patients were retrieved from the E-MTAB-1980 [43] and CPTAC [44] cohorts, respectively. The open-source MAAS is available from the following GitHub repository: https://github.com/Larrycpan/MAAS [128].

Acknowledgments

We thank Dr. Zheng Hu at the Chinese Academy of Sciences and members of the Li Laboratory for their helpful discussions. We also thank Dr. Zhiqiang Ye at the Shenzhen Bay Laboratory Supercomputing Center for high-level computing support.

Author contributions

L.L. conceived and supervised the project. K.X., R.D., Y.Q. and D.L performed the bioinformatics analysis. W.W. performed the experiments. J.W. and C.Y. contributed to the cancer analysis. K.X., R.D., X.Z., and L.L. wrote the manuscript with assistance from the other authors. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Supplementary information

Additional file 1. Description, tables and figures of additional results and methods. Fig. S1: Benchmarking CNV calling from scATAC-seq with DNA-seq data. Fig. S2: Scatter plots showing Spearman correlation between the copy number values measured from DNA sequencing and pseudo-bulk copy numbers estimated from scATAC-seq. Fig. S3: Evaluation of SNV detection from scATAC-seq against scWGS as the orthogonal ground truth. Fig. S4: Hamming distance between recovered SNV profile and ground truth. Fig. S5: Performances of cell clustering estimated by ARI, NMI and V-measure. Fig. S6: Accuracy of SNV denoising in two real colorectal cancer single-cell datasets. Fig. S7: MAAS performance with and without the chromatin accessibility correction. Fig. S8: MAAS performance against distinct value of CNV-weighting factor. Fig. S9: Contribution of each modality against distinct value of CNV-weighting factor. Fig. S10: MAAS performance against distinct convergence thresholds. Fig. S11: S-score against different dimensions of latent factor W and the number of clusters across datasets. Fig. S12: Comparing performances of tumor subpopulation identification by MAAS and other state-of-the-art methods. Fig. S13: Clustering stability metrics as a function of cell number. Fig. S14: Comparison of trees between ground truth and constructed by latent factors calculated by different methods estimated by mutual cluster information. Fig. S15: Clustering performance against different number of modalities in simulated data. Fig. S16: CPU time and maximum memory usage for each method against the total number of cells. Fig. S17: Identification of K562 cell subpopulations from paired scATAC-seq and scWGS data obtained from the same cells. Fig. S18: Identification of gastric cancer subpopulations using scDNA-seq and MAAS. Fig. S19: Clustering performance against

different number of modalities in real ovarian cancer dataset. Fig. S20: Identification of tumor cell subpopulations in treated and non-treated ovarian cancer samples. Fig. S21: Cell type annotation of primary pediatric posterior fossa ependymoma. Fig. S22: SNV profile of tumor cells of PPFE. Fig. S23: Tumor cell clusters defined by CNVs and chromatin states of pediatric ependymoma. Fig. S24: Correlation bewteen sequencing depth and variant allele frequency estimated by bootstrapping. Fig. S25: Driver gene mutation of *CASP9* was enriched in cluster 3. Fig. 26: Tumor cell clusters identified by traditional single-modality methods. Fig. S27: MAAS identified new B-cell lymphoma cell subpopulations. Fig. S28: SNV profile of tumor cells of glioma. Fig. S29: Copy number profiles of the glioma tumor cell clusters. Fig. S30: Tumor cell clusters defined by CNVs and chromatin states of glioma. Fig. S31: scATAC-seq peak tracks of clusters 1 and 2 for cluster 1-specific accessible regions. Fig. S32: DACRs between adult GBM clusters 1 and 2 validated on a pediatric GBM dataset. Fig. S33: Responses of MAAS-identified clusters temozolomide. Fig. S34: Responses of MAAS-identified clusters temozolomide in the validation dataset. Fig. S35: The TMZ-resisatance of MAAS clusters was independent of *MGMT* status. Fig. S36: Experimental validation of TMZ resistance of cluster 2 by examining cluster 2-specific genes. Fig. S37: Prognostic prediction performance. Fig. S38: C-index and AUC comparison for existing signatures across different cancer types. Fig. S39: Cox proportional-hazards model analysis revealed the prognostic value of MAASig in pan-cancer. Fig. S40: Calibration plot of the MAASig for survival outcomes. Fig. S41: Decision curve analysis of the MAASig for survival outcomes. Fig. S42: Tumor cell clusters defined by CNVs and chromatin states of renal cancer. Fig. S43: Identification of gene modules of each MAAS-determined cluster by

WGCNA. Table S1: Quality control metrics of scATAC-seq datasets analyzed in this study. Table S2: PCR primer. Table S3: shRNA oligonucleotides. Table S4: Datasets of MAASig performance estimation. Table S6: TFs and potential target genes included in the MAASig of each cancer. Table S6. Hazard ratio of signature genes across cancer types. Table S7: Publicly available signatures used for benchmark.

Additional file 2. Supplementary data. Data S1: List of genes in the drug-target gene module. Data S2: List of PPFE cluster 1-specific TFs and kinases included in the LINCS consortium. Data S3: List of differential chromatin accessible regions of glioma cluster 2 identified by MAAS.

Figures and Tables

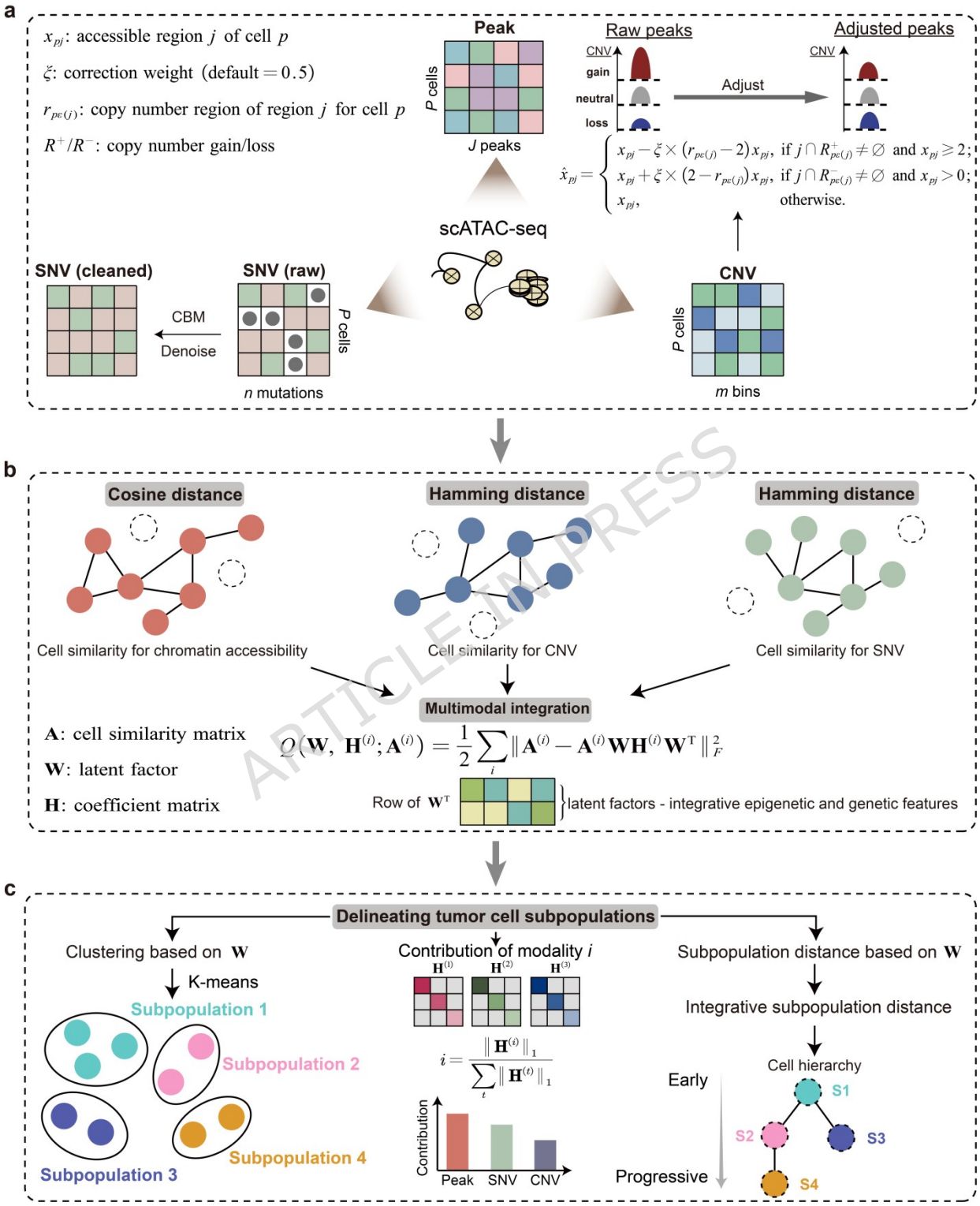

Fig. 1. **The MAAS workflow**. (**A**) MAAS takes as input a cell-by-peak matrix, a cell-by-CNV matrix, and a cell-by-SNV matrix. Raw peak data are adjusted based on the copy number values of the corresponding genomic regions. A robust principal component analysis (PCA) is applied to the SNV data to reduce noise and generate a low-rank matrix. (**B**). Cell similarities for each omics layer are calculated using Euclidean or Hamming distances. These similarities are integrated through a modified matrix factorization strategy, enabling the inference of a latent space that captures both genetic and epigenetic features through iterative updates. (**C**). Tumor cell subpopulations are identified using the latent factors. The contribution of each modality to the subpopulation is determined by the first-order norm of the coefficient matrix **H**. Consensus cell distances are derived by calculating the Euclidean distance from the cell-by-latent factor matrix, which is then used to reconstruct cell hierarchy represented by a minimum evolution tree.

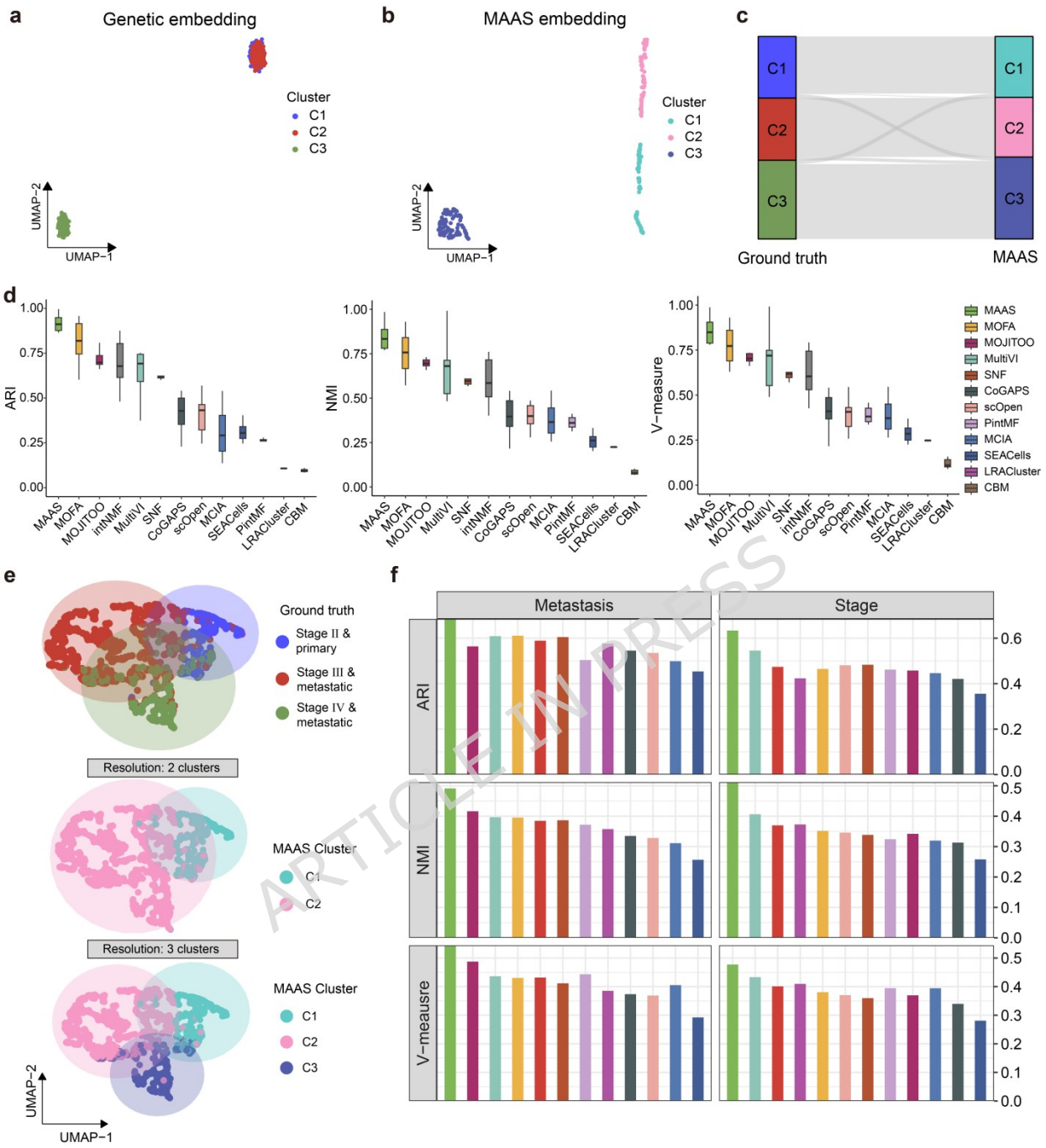

Fig. 2. Benchmarking analysis of tumor subpopulation identification. **a.** UMAP embedding based on genetic features of three subpopulations. **b.** UMAP embedding based on MAAS latent factors for the three identified subpopulations. **c.** Consistency of cell distribution across the three subpopulations when comparing ground-truth to MAAS results. **d.** Consistency of cell distributions across four subpopulations between the ground-truth and MAAS results. **e.** UMAP

embedding based on MAAS latent factors showing three ovarian cancer samples with different metastatic status and histological grade, and MAAS-identified clusters with distinct resolution. **f.** Accuracy of classifying tumor cell subpopulations across different computational methods. Classification performance is evaluated based on the ability to distinguish tumor subpopulations defined by metastatic status (primary vs. metastatic tumors, left panel) or pathological stages (II, III and IV, right panel).

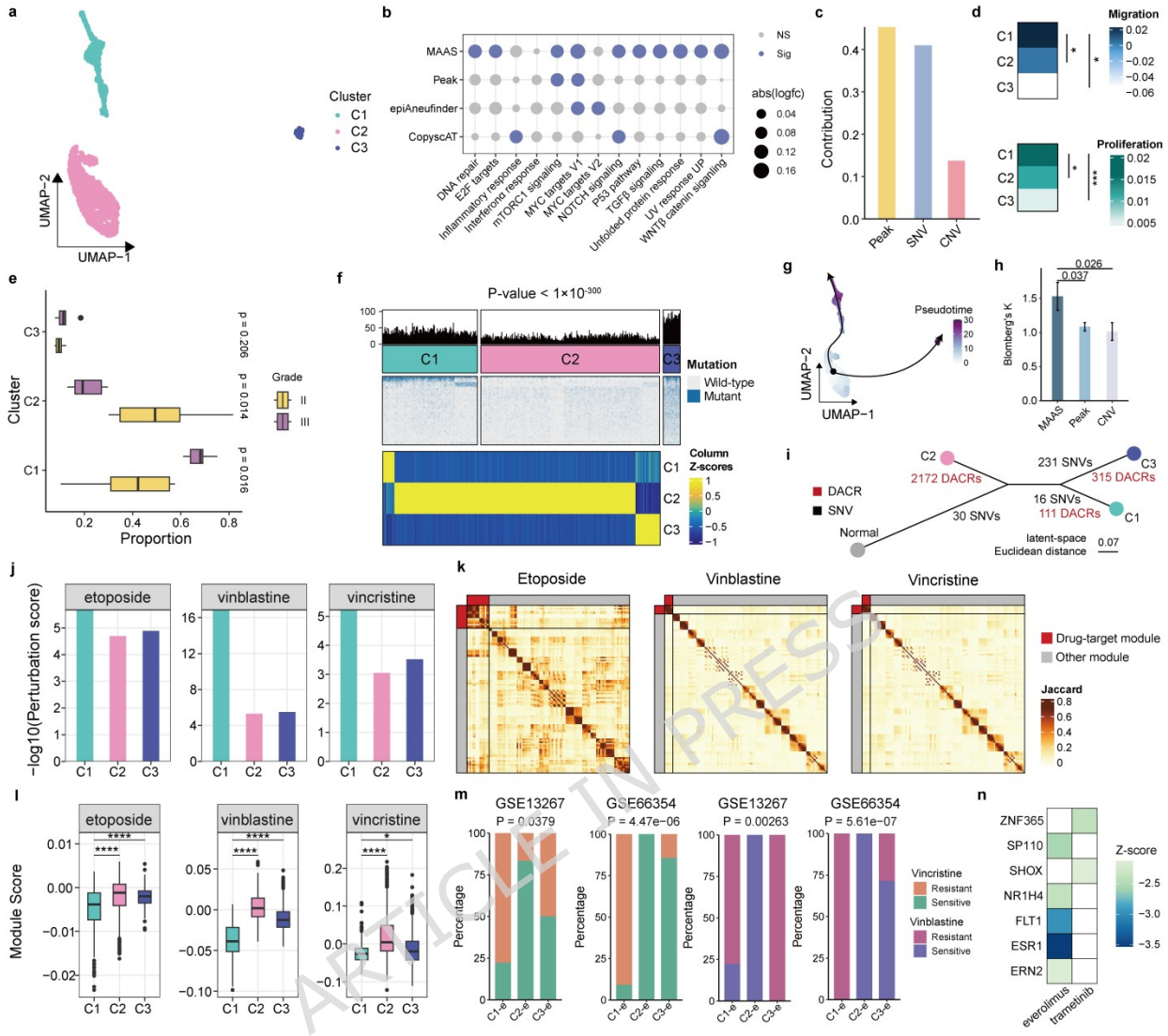

Fig. 3. A new pediatric ependymoma cell subpopulation with low CNV burden is associated with multidrug resistance. a. UMAP visualization of the two tumor cell subpopulations identified by MAAS. **b.** Significantly enriched cancer hallmarks in MAAS-identified clusters 1 and 3, with thresholds of $|\log FC| > 0.1$ and false discovery rate (FDR) < 0.05. **c.** Contribution of each modality to subpopulation identification. **d.** Average proliferation and migration scores for clusters 1 and 2. **e.** Cluster abundance between patients in grade II and III. *P*-values were determined by the t-test. **f.** Distribution of SNVs (top) and differentially accessible chromatin regions (bottom) across the three clusters. The heatmap in the bottom panel

shows Z-scored normalized accessibility. The *P*-value for mutational frequency differences was determined by the Kruskal-Wallis test. **g.** Pseudo-time ordering of tumor cell subpopulations showing their developmental trajectories. **h.** Blomberg's K values quantify evolutionary signals based on MAAS and single modalities. **i.** Evolution tree of MAAS-identified clusters depicting the temporal ordering of SNVs and DACRs. Edges represent Euclidean distances computed in the MAAS-derived latent space, with a unit branch length of 0.07. **j.** Perturbation scores for three first-line drugs, with lower scores indicating greater drug resistance. **k.** Gene modules inferred from drug target co-expression networks using non-negative matrix factorization. Low module scores indicate reduced predicted drug responsiveness. **l.** Boxplots comparing drug-target module scores between clusters 1 and 2. The center line represents the median, and the lower and upper hinges represent the first and third quartiles. The whiskers extend to the maximum and minimum values within 1.5 times the interquartile range from the hinge. *P*-values were determined by the Wilcoxon rank-sum test. **m.** Distribution of predicted drug-sensitive versus drug-resistant cells across MAAS-determined clusters, based on different bulk RNA-seq reference datasets. Statistical significance was assessed using Fisher's exact test. **n.** Heatmap showing reduced expression of upregulated transcription factors and kinases in cluster 1 following treatment with approved targeted therapies.

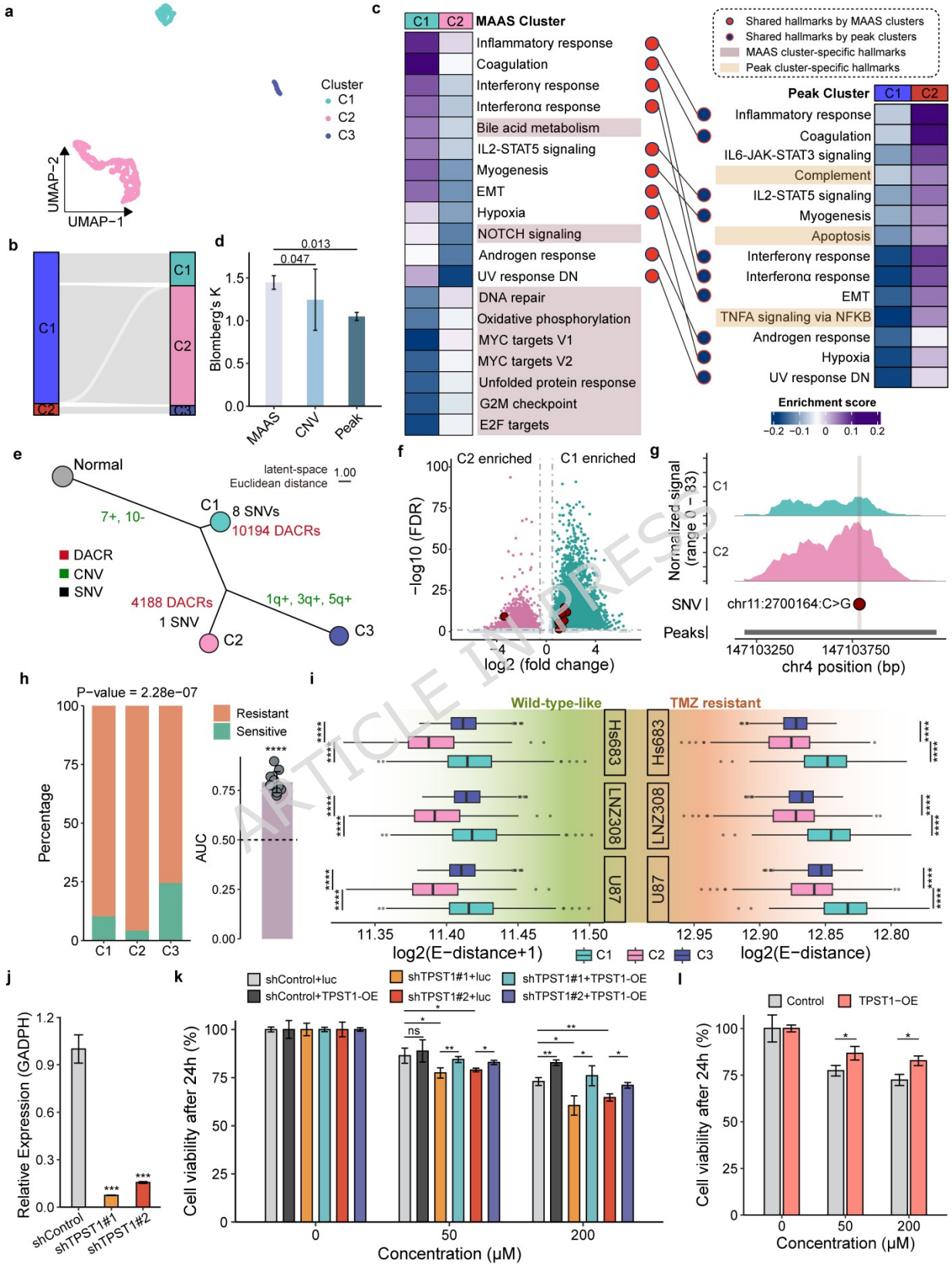

Fig. 4. High-resolution profiling reveals a TMZ-resistant glioma subpopulation. a. UMAP embedding of the three tumor cell clusters determined by MAAS. **b.** Sankey plot showed the correspondence between clusters identified by CNVs and those identified by MAAS. **c.** Differentially enriched cancer hallmark signatures between clusters. The left panel displays pathways enriched in MAAS clusters, while the right panel shows pathways enriched in tumor cell clusters identified by chromatin accessibility. The heatmap colors represent enrichment scores, with color shades indicating unique hallmarks detected by either MAAS or chromatin accessibility. **d.** Blomberg's K quantifies the evolution signal of tumor cell clusters by MAAS and traditional single-modality methods. **e.** Hierarchy of MAAS-identified clusters, depicting the timing of CNVs, SNVs, and DACRs. Edges represent Euclidean distances computed in the MAAS-derived latent space, with a unit branch length of 1.00. **f.** Volcano plot showing DACRs (FDR < 0.05 and |logFC| > 0.5) between MAAS-identified clusters 1 and 2. Regions containing the six driver mutations specific to cluster 1 are highlighted. **g.** scATAC-seq peak tracks for accessible regions in clusters 1 and 2, with noncoding SNVs marked by dark red dots. **h.** Distribution of predicted temozolomide (TMZ)-sensitive and resistant cells across MAAS-identified clusters. The *P*-value, calculated by a chi-square test, is shown, along with the accuracy of predictions measured by the area under the curve (AUC). The dotted line represents the baseline of 0.5. Asterisks indicate significance based on permutations. **i.** Energy distance (E-distance) between the three MAAS clusters and TMZ-sensitive and resistant cells across three cell lines. The center line of the boxplot indicates the median, the box limits show the first and third quartiles, and the whiskers extend to the maximum and minimum values within 1.5 times the interquartile range from the hinge. The *P*-value was determined by a two-tailed Wilcoxon rank-sum test. **j.** *TPST1* gene expression measured by qPCR. **k.** Effect of *TPST1* knockdown and

overexpression on TMZ resistance across different concentrations. The viability of cells was measured after treatment with various concentrations of TMZ (0, 50, and 200 μM) in different experimental groups. Groups include: shControl+luc (shRNA non-targeting and luciferase), sh*TPST1*#1+luc and sh*TPST1*#2+luc (*TPST1* knockdown with luciferase control), and shControl+*TPST1*-OE, sh*TPST1*#1+*TPST1*-OE, and sh*TPST1*#2+*TPST1*-OE (*TPST1* overexpression). Viability is shown as the mean ± standard error of the mean across biological replicates (n = 3). Statistical significance was assessed using a Student's t-test. **l.** Cell viability of *TPST1* and overexpression between control and TMZ-treatment groups. *P*-values were determined by Student's t-test. ns: no significance, **P*-values < 0.05, ***P*-values < 0.01, ****P*-values < 0.001, *****P*-values < 0.0001.

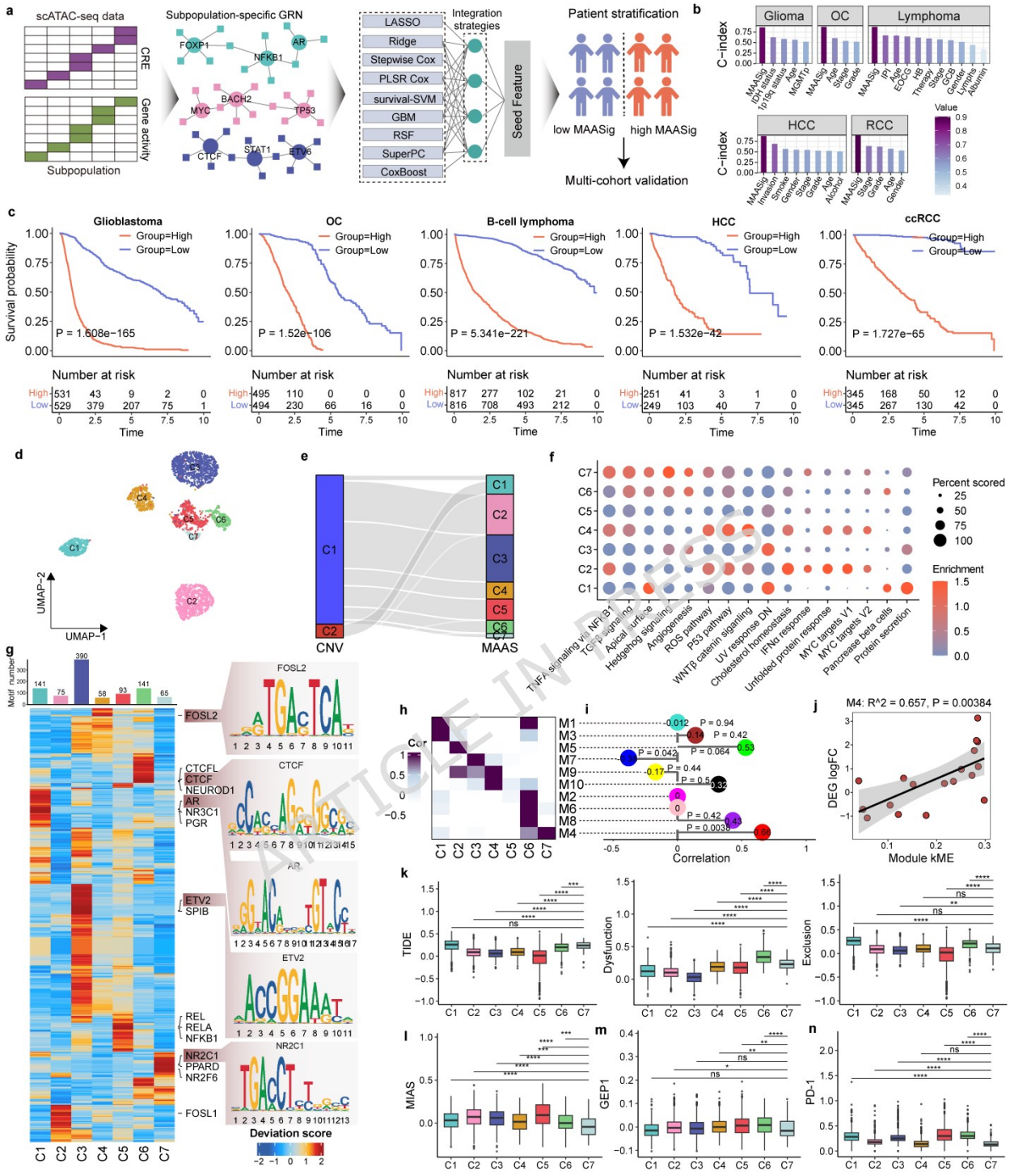

Fig. 5. A MAAS-derived clinical signature accurately predict prognosis across multiple cancer types. a. Schematic illustration of the workflow for generating the MAAS-derived multimodal signature (MAASig) (see details in the Supplementary Materials and Methods). **b.**

Average concordance index (C-index) of traditional clinical features and MAASig for survival prediction across multiple cancer types, including glioblastoma, ovarian cancer (OC), B-cell lymphoma, hepatocellular carcinoma (HCC), and clear cell renal cell carcinoma (ccRCC). **c.** Kaplan–Meier survival curves demonstrating the clinical relevance of MAASig in a pan-cancer meta-cohort. Datasets for each cancer type were combined into a single cohort, with MAASig stratification determined at the median value. Statistical *P*-values were calculated using a two-tailed log-rank test. **d.** UMAP embedding of the six tumor cell subpopulations identified by MAAS. **e.** Sankey plot illustrating the clusters identified by CNVs and those identified by MAAS. **f.** Cancer hallmark pathways enriched in each cluster. **g.** Heatmap of chromVAR bias-corrected deviation scores for the differential TF motifs across clusters. The top bar indicates cluster-specific TF motifs with examples of sequence logos for the top TF motifs displayed on the right side of the plot. **h-j.** Spearman correlations: between eigengene-based connectivity (kME) of all modules (**h**), between module 4 eigengene and kME (**i**), and between log fold change (logFC) of differentially expressed genes and anti-PD-1 response versus non-response in patients (**j**). Shaded areas represent 95% confidence intervals. **k-n.** Degree of immunotherapeutic response measured by various metrics: tumor immune dysfunction (**k**) and exclusion (TIDE) scores, MHC I association immunoscore (MIAS) (**l**), 18-gene expression profiles (GEP) (**m**), and PDCD1 (PD-1) (**n**) gene activity. The center line in each box plot represents the median, the lower and upper hinges represent the first and third quartiles, and the whiskers extend to the maximum and minimum values within 1.5 times the interquartile range from the hinge. *P*-values were determined using the Wilcoxon rank-sum test. ns: no significance, **P*-values < 0.05, ***P*-values < 0.01, ****P*-values < 0.001, *****P*-values < 0.0001.