Supplementary Material 1

Detailed protocols for laser microdissection preparation and bioinformatics analysis

Laser microdissection-based multi-omics integration unveils the pathological atlas and tumor differentiation network of adamantinomatous craniopharyngioma

Bowen Wu[1] #, MM, Chunming Xu[1] #, MD, Chenxing Yang[1] #, MM, Shenhao Xie[1] #, MM, Suping Duan[2], BMed, Jiye Ye[3], MD, Jinshi Zhang[4], MD, Shaoyang Li[5], MD, Furong Liang[1], BMed, Zhicheng Ye[1], BMed, Hai Luo[1], MM, Rilege Su[1], BMed, Laisheng Pan[1], MM, Zhizheng Liu[1], MM, Shizhou Xing[1], BMed, Tianming Dong[1], BMed, Yiguang Chen[6], MD, Like Chen[1], MM, Suyue Zheng[1], MM, Jie Wu[1] *, MD, Tao Hong[1] *, MD

[1] Department of Neurosurgery, Jiangxi Key Laboratory of Neurological Diseases, the First Affiliated Hospital, Jiangxi Medical College, Nanchang University, Nanchang, 330006, China

[2] Department of Neurology, the First Affiliated Hospital, Jiangxi Medical College, Nanchang University, Nanchang, 330006, China

[3] Department of Neurosurgery, Hunan University of Medicine General Hospital, 418000, Huaihua, China

[4] Department of Neurosurgery, Ganzhou Hospital-Nanfang Hospital, Southern Medical University (Ganzhou People's Hospital), Ganzhou, 341000, China

[5] School of public health, Jiangxi medical college, Nanchang University, Nanchang, 330006, China

[6] Department of Neurosurgery/Neuro-oncology, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, SunYat-sen University Cancer Center, 510060, Guangzhou, China

# Bowen Wu, Chunming Xu, Chenxing Yang and Shenhao Xie contributed equally to this work.

* Corresponding author:

Jie Wu: Department of Neurosurgery, Jiangxi Key Laboratory of Neurological Diseases, the First Affiliated Hospital, Jiangxi Medical College, Nanchang University, 17 Yong Wai Zheng Street, Nanchang, 330006, China, E-mail: ndyfy10128@ncu.edu.cn

Tao Hong: Department of Neurosurgery, Jiangxi Key Laboratory of Neurological Diseases, the First Affiliated Hospital, Jiangxi Medical College, Nanchang University, 17 Yong Wai Zheng Street, Nanchang, 330006, China, E-mail: ndyfy00567@ncu.edu.cn

Preparation of materials and slides before laser microdissection

All experimental instruments and materials were disinfected by autoclaving or ultraviolet irradiation, and the work surface was wiped with RNase solution. After the samples were removed from liquid nitrogen, they were quickly embedded in precooled OCT compound and allowed to solidify in the chamber of the Leica cryostat (-20°C). Continuous sections were cut at a thickness of 10 μm and fixed on 4-μm PEN membrane slides (Leica 11600288, presoaked in absolute ethanol for 30 minutes and naturally air-dried). The tissue sections were subsequently treated with 95% ethanol, followed by hematoxylin staining, differentiation, reverse blue staining, eosin staining, and gradient ethanol dehydration. All steps involved intermittent rinsing with DEPC-treated water. Before the subsequent experimental steps, the slides were naturally air-dried. Fresh reagents were used for each experiment, and all water solutions were reconfigured using newly opened DEPC-treated water.

Batch effect removal and data quality control

With respect to the Smart-Seq2 data, to eliminate the influence of batch effects on gene expression levels, we applied the ComBat function from the R software package sva to adjust the original count, TPM, and FPKM expression matrices. During the adjustment process, the pathological structure was retained as a known biological covariate, and the experimental batch was treated as a nuisance covariate for removal.

For the single-nuclear RNA sequencing (snRNA-seq) data, the analyses were performed using the Seurat R package[1]. We first used DoubletFinder[2] to calculate and remove double cells for each individual sample and then merged the matrices and set thresholds consisting of the UMI count, number of expressed genes, and percentage of mitochondrial gene count to remove low-quality cell nuclei. We subsequently standardized the data using the SCTransform method and corrected the mitochondrial gene content and cell cycle effects. Then, based on 3000 highly variable genes, we integrated multiple samples to eliminate batch effects. We performed PCA linear dimensionality reduction and selected the first 30 principal components based on the elbow rule for UMAP visualization and cell clustering. Finally, we conducted unsupervised clustering using the FindClusters function and identified cell subpopulations using cell type marker genes.

For the ST data, we performed strict data quality control for each sample independently and used SCTransform for standardization. Based on the principal component analysis for dimensionality

reduction, and followed by the selection of an appropriate resolution for clustering, the spot clusters were finally mapped back to the corresponding hematoxylin and eosin (H&E)-stained images according to their spatial coordinates. Spots were defined by observing the characteristic histopathological morphology.

Similarity analysis

To explore the similarities in the transcriptomes between different samples or pathological structures, we performed a correlation analysis. For the Smart-Seq2 data, we calculated the correlation coefficient matrix for all tumor samples pairwise. This analysis was based on the $\log_2(TPM + 1)$ expression matrix after batch correction, covering all expressed genes.

For snRNA-seq and ST data, to make comparisons at the population level, we first calculated the average gene expression profiles of each cell cluster. We subsequently selected the 1000 genes with the greatest expression variability across all cell clusters and constructed a Spearman correlation coefficient matrix for pairwise comparisons of cell clusters based on this gene set.

Spatial colocalization analysis

Additionally, based on the proportion data of pathological features obtained from robust cell type decomposition analysis[3], we further evaluated the spatial colocalization between paired cell types by using the bivariate Moran's I analysis. For each pair of cell types, we extracted the feature scores corresponding to all spatial points and constructed a spatial weight matrix based on the spatial distance to define the neighborhood relationship. The statistical significance of Moran's I was evaluated through permutation tests. All analyses were completed using the MERINGUE R package[4]. A positive and significant Moran's I ($I_{AB} > 0$, $p < 0.05$) indicated spatial colocalization between the two cell types.

Differential gene expression and enrichment analysis

With respect to the Smart-Seq2 data, based on the batch-corrected combat-count matrix, we used the DESeq2[5] and edgeR[6,7] packages to conduct pairwise screening of differentially expressed genes (DEGs). For the snRNA-seq and spatial transcriptomic (ST) data, the FindAllMarkers function in Seurat was used to identify the DEGs in each cell cluster that were different from those in other clusters. All methods used a filtering threshold of a log2FC absolute value greater than or equal to 1 and an

adjusted *p* value (padj) less than 0.05.

We performed functional enrichment analysis based on the list of differentially expressed upregulated genes using the screened DEGs and conducted gene set enrichment analysis (GSEA) using the log2FC values of all genes as the sorting indicator. Both analyses were conducted for the signature gene sets of MSigDB and were implemented by the clusterProfiler R package[8,9]. Additionally, we submitted the differentially expressed upregulated genes of specific subgroups to the Metascape online platform for comprehensive analyses, including transcription factor target enrichment and protein–protein interaction (PPI) network construction[10].

Identification of high-confidence differentially expressed genes through multiple comparative intersection

With respect to the Smart-Seq2 data, to identify more conservative and stable gene expression patterns and biological processes with specific pathological structures, we performed intergroup comparisons between structures (such as WC vs. PE, WC vs. SR, and WC vs. CE). The resulting lists of these DEGs were subjected to a multiple intersection procedure to extract a high-confidence set of DEGs, which were subsequently used for functional enrichment analysis.

Gene set variation analysis

We downloaded five core gene sets, namely, KEGG, GO, WikiPathways, Reactome, and Hallmark, from the MSigDB database and processed them into the required format for analysis using the msigdbr R package. For the data obtained from the snRNA-seq, we used the average expression function in the Seurat package, taking the data from the RNA assay as input, to calculate the average expression levels of genes within each cell cluster of different pathological structures, after which the results were used as the input matrix for gene set variation analysis (GSVA). The GSVA R package was subsequently used to calculate the pathway enrichment scores for gene sets[11], and the resulting scores were visualized using the pheatmap package to demonstrate the differences in pathway activity among different pathological structures.

Weighted gene coexpression network analysis

We performed weighted gene co-expression network analysis (WGCNA) using the batch-

corrected $\log_2$(FPKM + 1) expression matrix from Smart-Seq2[12]. We constructed a network through an automatic one-step method and identified gene modules with highly coordinated expression patterns among samples. Each gene module has a set of characteristic genes, and we subsequently screened out the modules significantly related to key biological traits by calculating the correlation between the modules and pathological structures. To elucidate the biological functions of these co-expression modules, we performed GO, KEGG, WikiPathways, and Reactome pathway enrichment analyses on all genes within each significant module as inputs to systematically reveal the potential biological processes, molecular functions, and signaling pathways represented by the modules.

Score of functional gene sets

For the data obtained by Smart-Seq2, the $\log_2$(TPM + 1) transformation and ComBat batch-corrected expression matrix were used as inputs. The gene set enrichment scores were calculated separately using the GSVA and ssGSEA algorithms (with kcdf = "Gaussian"). We subsequently performed hierarchical clustering and constructed heatmaps based on the obtained enrichment score matrices according to the experimental groups. With respect to the snRNA-seq data, to evaluate the activity of specific gene sets in different pathological features, we used the AddModuleScore function in the Seurat package to calculate the functional scores of each cell, after which the differences in this functional score among different pathological structures were visualized and compared through violin plots.

Effect sizes of AddModuleScore analysis

Pairwise comparisons were performed using the Wilcoxon rank-sum test, and *p* values were calculated via Monte Carlo simulation with 10,000 replicates. Effect size r was derived from the standardized statistic Z using the formula:

$$r = \frac{Z}{\sqrt{N}}$$

N is the total number of observations in the two groups. *p* values were adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) method.

Cell-cell communication

We used CellPhoneDB (v4.0.0) to infer the communication between cell subpopulations[13]. By using Scissor-mapped pathological feature subpopulations and the standardized expression matrix as the input, we conducted inference following the standard statistical analysis methods of the tool. The significant intercellular ligand–receptor interactions were visualized by the ktplots R package.

Copy number variation analysis

To identify copy number variations (CNVs) across different pathologies, we used the inferCNV algorithm with various lymphoid and myeloid immune cells (or pituitary cells) with normal diploid characteristics as references to analyze the snRNA-seq data. We systematically inferred the chromosomal copy number abnormalities present in each epithelial subset.

Transcription factor analysis

In the Python environment, we used the decoupleR package to infer the specific transcription factor activity of each cell cluster[14]. We constructed an experimentally validated transcription factor–target gene regulatory network from the COLLECTRI database. We subsequently estimated the activity score of each transcription factor in each cell using the ulm algorithm based on the single-cell gene expression matrix. Finally, through the analysis of intergroup differential activity, we identified the significantly activated transcription factors in each cell cluster and compared their activity and mRNA expression patterns through multipanel visualization. For each transcription factor, we generated a combined graph, including its activity and expression distribution in the UMAP space and the activity and expression violin plots in each cell cluster.

Temporal analysis

To systematically analyze the differentiation dynamics of pathological epithelial cell subpopulations, we integrated multiple computational methods. First, on the basis of the original RNA count matrix, the CytoTRACE algorithm was used to assess the developmental potential of the Scissor-mapped pathological subpopulations, and the degree of developmental potential of each subpopulation was compared through box plots[15].

On this basis, we used Monocle2 to infer cell trajectories, using the highly variable genes selected

by SCTransform as the characteristic gene set to construct a pseudotime developmental path[16]. Based on the results of CytoTRACE, we used the high-potential progenitor cell subpopulations as the root point of the trajectory and further conducted GO enrichment analysis on the differentially expressed genes in the branches to reveal the key biological processes involved in cell fate differentiation.

Finally, to explore in depth the coordinated regulatory networks during state transitions, we jointly applied the Monocle3[17] and high-dimensional WGCNA (hdWGCNA) methods[18], using the pseudotime generated by Monocle3 as the phenotypic input to construct a weighted gene coexpression network, identifying the gene modules that were specifically activated during evolution from the root subpopulation to other pathological features. Through functional enrichment and visualization of the pseudotime expression of genes in this module, we systematically elucidated the core transcriptional program driving the early pathological structural differentiation transformation of epithelial cells.

RNA velocity analysis

To elucidate the differentiation dynamics among pathological structures, we performed an RNA velocity analysis based on snRNA-seq data[19]. By processing the BAM files generated by Cell Ranger using Velocyto, we obtained information on the unspliced and spliced transcriptomes of genes. The scVelo tool was subsequently used to infer the future state of each cell, and the calculated rate vectors were projected onto the UMAP dimensionality-reduced space in the form of a flow field, thereby visualizing the potential direction of cell differentiation.

References:

1. Hao Y, Hao S, Andersen-Nissen E et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573-3587.e29
2. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst*. 2019;8(4):329-337.e4
3. Cable DM, Murray E, Zou LS et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol*. 2022;40(4):517-526
4. Miller BF, Bambah-Mukku D, Dulac C, Zhuang X, Fan J. Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities. *Genome Res*. 2021;31(10):1843-1855
5. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550-550
6. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*. 2010;26(1):139-140
7. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*. 2012;40(10):4288-4297
8. Yu G, Wang L, Han Y, He Q. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology*. 2012;16(5):284-287
9. Wu T, Hu E, Xu S et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (N Y)*. 2021;2(3):100141-100141
10. Zhou Y, Zhou B, Pache L et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun*. 2019;10(1):1523-1523
11. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*. 2013;14:7-7
12. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9:559-559
13. Garcia-Alonso L, Lorenzi V, Mazzeo CI et al. Single-cell roadmap of human gonadal development. *Nature*. 2022;607(7919):540-547
14. Badia-I-Mompel P, Vélez Santiago J, Braunger J et al. decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics advances*. 2022;2(1):vbac016-vbac016
15. Gulati GS, Sikandar SS, Wesche DJ et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science (New York, N.Y.)*. 2020;367(6476):405-411
16. Qiu X, Mao Q, Tang Y et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods*. 2017;14(10):979-982
17. Cao J, Spielmann M, Qiu X et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*. 2019;566(7745):496-502
18. 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(6):100498-100498
19. Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol*. 2020;38(12):1408-1414