Articles

DOI:https://dx.doi.org/10.71373/KKRB6502
Submitted 25 October 2025
Accepted 22 December 2025
Pulished 25 December 2025

Revealing Causal Genes in Nasopharyngeal Carcinoma: An Integrated Approach of Mendelian Randomization and Single-Cell Sequencing

Yuxiao He[1#] , Bin Han[1#] , Xuanyu Pan[2] , Jianping Zhou[1] , Tingwei Lv[1] , Xingrao Li[3] , Yongjian Zhou[1] , Yani Yang[3] , Qingniao Zhou[2] , Tingting Liu[4] ,Yanling Hu[1,2,3*]

Objective:Nasopharyngeal carcinoma (NPC) is a malignancy with a complex genetic basis and a highly heterogeneous tumor microenvironment. This study aimed to identify genes causally associated with NPC risk using Mendelian randomization (MR) and to elucidate their cell-type-specific expression patterns and disease relevance by integrating single-cell RNA sequencing (scRNA-seq) data. Methods: Bulk transcriptomic datasets of NPC and non-tumor nasopharyngeal tissues were obtained from the Gene Expression Omnibus (GEO) to identify shared differentially expressed genes (DEGs). Expression quantitative trait loci (eQTLs) corresponding to these genes were retrieved from OpenGWAS, and NPC genome-wide association study (GWAS) summary statistics were obtained from the FinnGen database. Gene expression levels, represented by eQTLs, were used as exposures in MR analyses, including Wald ratio, inverse-variance weighted (IVW), MR-Egger, and weighted median methods. To link genetic findings to disease biology, scRNA-seq datasets of NPC and non-tumor samples were integrated, batch-corrected, and annotated using a standard analysis workflow. Cell-type-specific expression patterns and tumor–non-tumor expression differences of MR-identified genes were systematically evaluated. Results: A total of 494 shared DEGs were identified, among which 313 genes were eligible for MR analysis. MR analyses identified five genes with putative causal associations with NPC risk: *TMEM200A* and *THBS2* as protective factors, and *MFSD4*, *PSPH*, and *VPREB3* as risk factors.Single-cell analysis revealed distinct cell-type-specific expression patterns of these genes within the NPC tumor microenvironment. Notably, *TMEM200A* was enriched in fibroblasts and T cells, *THBS2* in fibroblasts, *MFSD4* in endothelial cells, *PSPH* in epithelial cells, and *VPREB3* in B cells and plasma cells. Furthermore, several of these genes exhibited significant expression differences between tumor and non-tumor samples within specific cell populations, providing direct cellular evidence linking MR-identified genetic risk signals to NPC-related biological processes. Conclusions: By integrating Mendelian randomization with single-cell transcriptomic analysis, this study bridges genetic causal inference with cell-type-resolved disease biology in NPC. The identified genes exhibit distinct cellular localizations and tumor-associated expression alterations, highlighting the contributions of immune, stromal, endothelial, and epithelial compartments to NPC susceptibility. These findings provide a refined framework for understanding NPC pathogenesis and offer potential targets for future functional and therapeutic studies.

Introduction

Nasopharyngeal carcinoma (NPC) is an epithelial malignancy characterized by distinct geographic and ethnic distribution patterns, with a notably high incidence in regions such as southern China and Southeast Asia, where it poses a considerable public health challenge[1]. The early clinical manifestations of NPC are often insidious and nonspecific, resulting in frequent diagnosis at advanced stages and generally unfavorable prognosis. A deeper understanding of NPC pathoge0 nesis is therefore essential to inform early detection and the development of effective therapeutic strategies.

Mendelian randomization (MR) has emerged as a powerful methodological framework for inferring causal relationships between exposures and disease outcomes. By leveraging genetic variants as instrumental variables, MR minimizes confounding and reverse causation biases that commonly affect conventional observational studies[2]. While MR has been applied to identify environmental influences on NPC risk — such as allergic rhinitis[3] and vitamin D levels[4]—a systematic investigation into the causal roles of genetic determinants in NPC remains limited.

To address this research gap, we integrated transcriptomic and genetic data to systematically identify and validate candidate causal genes for NPC. Our analytical approach consisted of three main stages: first, we identified differentially expressed genes from bulk RNA sequencing data; second, we performed MR analysis using large-scale GWAS summary statistics to evaluate their potential causal effects on NPC; and finally, we employed single-cell RNA sequencing to resolve the expression patterns of these genes across cell types within the tumor microenvironment. This multi-faceted strategy provides new mechanistic insights into NPC etiology and highlights potential targets for future gene-directed therapies.

Data and Methods

Transcriptomic Data Analysis

Two nasopharyngeal carcinoma (NPC) bulk transcriptomic datasets, GSE53819 and GSE12452, were retrieved from the Gene Expression Omnibus (GEO) database, comprising a total of 49 NPC tissues and 28 non-cancerous nasopharyngeal tissues. Raw expression data were downloaded from their respective platforms, and probe identifiers were mapped to standardized gene symbols using R (version 4.4.1). Expression profiles within each dataset were integrated using the merge function, and genes with missing values were removed to construct independent expression matrices for subsequent analyses.

1. Institute of Life Sciences, Guangxi Medical University, Nanning, China , 530021; 2. School of Basic Medical Sciences, Guangxi Medical University, Nanning, China , 530021. 3. Guangxi Medical University School of Information and Management, Nanning, China , 530021. 4.University of The Thai Chamber of Commerce, Bangkok, Thailand, 10400

#Yuxiao He and Bin Han contributed equally to this study

**Corresponding author*

E-mail addresses: huyanling@gxmu.edu.cn (Yanling Hu)

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Yanling Hu

Differential expression analysis was performed using the "*limma*" package (version 3.60.6). The analysis was conducted on quantified gene expression matrices that had been normalized across samples during upstream preprocessing and were represented as $\log_2$(TPM + 1) values. As the input data consisted of log-transformed continuous expression values and no longer followed the mean–variance relationship characteristic of raw count data, the voom transformation was not applied.

Linear models were fitted using the lmFit function, and contrast matrices were constructed with makeContrasts to enable groupwise comparisons. Empirical Bayes moderation was subsequently applied using the eBayes function to improve the robustness of statistical inference. Multiple testing correction was performed using the Benjamini–Hochberg method, and genes with an adjusted P value < 0.05 were considered statistically significant. Differentially expressed genes (DEGs) were extracted using the topTable function, with entries containing missing values excluded. Volcano plots were generated using the HiOmics cloud platform (https://www.henbio.com/) [5].

Mendelian Randomization Analysis

The overlapping differentially expressed genes (DEGs) were converted to Ensembl gene identifiers using the Ensembl BioMart database. In the Mendelian randomization analysis, the differential expression status of genes was not directly used as the exposure variable, but served only to screen for genes with potential biological relevance. The actual exposure variable was defined as the expression level of the corresponding gene, represented by expression quantitative trait loci (eQTLs). For each candidate gene, summary statistics of associated eQTLs were obtained from the OpenGWAS database, and single nucleotide polymorphisms (SNPs) significantly associated with gene expression were selected as instrumental variables (IVs). IVs were filtered based on the significance of SNP-gene expression associations ($p \leq 1\times10^{-5}$). Selected SNPs then underwent linkage disequilibrium clumping (kb = 10,000, R^2 = 0.001) to ensure independence among instruments, and palindromic variants were excluded. The strength of the instrumental variables was assessed using the F-statistic, with F > 10 considered sufficient to avoid weak instrument bias[6]. Finally, using gene expression as the exposure and GWAS summary statistics for nasopharyngeal carcinoma from the FinnGen database as the outcome, Mendelian randomization analysis was performed to evaluate potential causal relationships between gene expression and NPC risk.

MR analyses were conducted under three core assumptions: (1) IVs are strongly associated with the exposure (gene expression); (2) IVs are independent of confounders; and (3) IVs affect the outcome (NPC) only through the exposure. The inverse variance weighted (IVW) method served as the primary approach[6]. Four supplementary methods were also applied: MR-Egger, weighted median, simple mode, and weighted mode[7-9].

Sensitivity analyses included Cochran's Q test to assess heterogeneity and MR-Egger intercept test to evaluate horizontal pleiotropy. A p-value < 0.05 indicated significant heterogeneity or potential pleiotropy[10, 11]. All analyses were performed in R using the "*TwoSampleMR*" (v0.6.8) and "*MR-PRESSO*" (v1.0) packages.

Single-Cell RNA-Seq Processing and Analysis

Two single-cell RNA sequencing datasets of nasopharyngeal carcinoma, GSE150825 and GSE120926, were downloaded from the Gene Expression Omnibus (GEO) database. GSE150825 comprises 3 samples of nasopharyngeal lymphoid hyperplasia and 11 nasopharyngeal carcinoma samples, while GSE120926 includes 16 nasopharyngeal carcinoma samples and 8 non-malignant nasopharyngeal tissue samples. All non-malignant samples and lymphoid hyperplasia samples were uniformly classified as the non-cancerous control group.

Single-cell data analysis was performed in the R environment using the "*Seurat*" package (version 4.1.1). Quality control is achieved by selecting cells that contain at least 200 detectable genes and a mitochondrial gene expression ratio below 15%. Subsequently, cell cycle stages were assigned to each cell using the "CellCycleScoring" function to assess potential cell cycle effects.

During data preprocessing, the gene expression matrix of each dataset was log-normalized using the "LogNormalize" method in Seurat (scale factor = 10,000). To capture the major biological variations across cells, highly variable genes were selected using the "FindVariableFeatures" function, retaining the top 2000 most variable genes. The data were then linearly scaled based on these selected genes using the "ScaleData" procedure, followed by principal component analysis (PCA).

To mitigate batch effects between the two datasets, the Harmony algorithm was applied to integrate them within the PCA space. Using the first 15 principal components after integration, cell-cell proximity was computed with the "FindNeighbors" function to construct a k-nearest neighbor graph for downstream analysis. UMAP dimensionality reduction was then performed for visualization and unsupervised clustering.

The choice of clustering resolution was evaluated using the "*Clustree*" package (version 0.5.0) by comparing cluster stability and hierarchical relationships across multiple resolution parameters. A resolution of 0.2 was ultimately selected. Differentially expressed genes for each cell cluster were then identified using the "FindAllMarkers" function.

Cell type annotation was performed manually based on the top 100 marker genes per cluster, combined with known canonical cell marker genes and relevant literature. Finally, the expression patterns of key genes identified through Mendelian randomization analysis across different cell subtypes were visualized using the "DotPlot" function in Seurat, where dot color and size represent the average expression level and the proportion of cells expressing the gene, respectively. In addition, stacked bar plots illustrating cellular composition between tumor and non-tumor tissues, as well as violin plots and boxplots showing the expression of key genes in corresponding cell types, were generated using the "*ggplot2*" package (version 3.3.6).

Results

Identification of Differentially Expressed Genes

Differential expression analysis was performed on the GSE53819 and GSE12452 datasets by comparing tumor tissues to non-cancerous controls, using thresholds of $|\log_2 FC| \geq 1$ and adjusted P < 0.05. In the GSE12452 dataset, 939 differentially expressed genes (DEGs) were identified, including 575 up-regulated and 364 down-regulated genes. The GSE53819 dataset yielded 4,212 DEGs, of which 1,823 were up-regulated and 2,389 were down-regulated (Figures 1A, B). A Venn diagram generated with the R package "*ggvenn*" (v0.1.10) revealed 494 shared DEGs between the two datasets (Figure 1C).

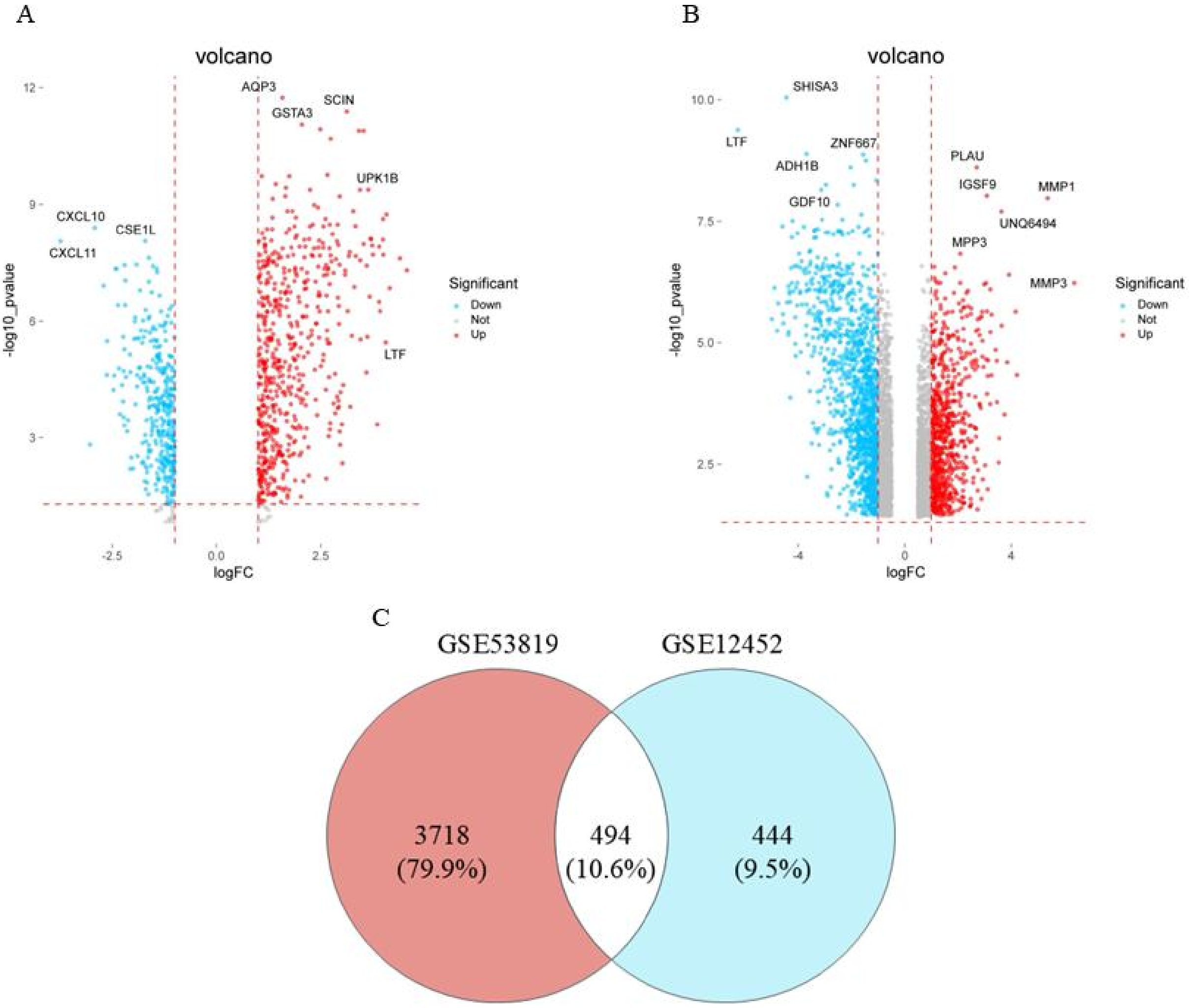

Figure1. (A)Volcano plot of differentially expressed genes between the healthy group and the tumor group in the GSE12452 dataset. (B)Volcano plot of differentially expressed genes between the healthy group and the tumor group in the GSE53819 dataset. (C)Venn diagram illustrating the shared differentially expressed genes across compared groups.

Mendelian Randomization Identifies Causal Genes for NPC

GWAS data for eQTLs of 313 shared DEGs were obtained from the OpenGWAS database, and summary statistics for nasopharyngeal carcinoma (FinnGen code: C3_NASOPHARYNX_EXALLC) were retrieved from the FinnGen consortium. MR analysis was conducted to estimate the causal effects of these genes on NPC risk. The Wald ratio method was applied for exposures with one SNP, and the inverse variance weighted (IVW) method was used for exposures with two or more SNPs. Four supplementary methods—MR-Egger, weighted median, simple mode, and weighted mode—were also applied for robustness. Associations with P < 0.05 in the primary method (Wald ratio or IVW) were considered statistically significant. Wald ratio analysis identified *MFSD4* as a risk gene for NPC (OR = 11.020, 95% CI: 1.637-74.150, P = 0.014)。And the wide confidence interval observed for *MFSD4* likely reflects the limited number of instrumental variables available for this gene and the relatively large standard error of the Wald ratio estimate.(Table1)

Table 1 Mendelian randomization results for *MFSD4*

Exposure	Snp	Methods	OR(95%CI)	Beta	P-value
MFSD4	1	Wald ratio	11.020 (1.637-74.150)	2.399	0.014

IVW analysis of exposures with exactly two SNPs revealed that *PSPH* was associated with increased NPC risk (OR = 3.942, 95% CI: 1.135-13.685, P = 0.031), whereas *THBS2* was associated with reduced risk (OR = 0.211, 95% CI: 0.047-0.939, P = 0.041). No significant heterogeneity was detected for these associations (Cochran's Q P > 0.05). (Table 2)

Table 2 Mendelian randomization results for *PSPH* and *THBS2*

Exposure	Snp	Methos	OR (95%CI)	Beta	P-value	Heterogeneity	
						Cochran'sQ	P-value
PSPH	2	IVW	3.942 (1.135-13.685)	1.372	0.031	0.360	0.548

Exposure	Snp	Methos	OR (95%CI)	Beta	P-value	Heterogeneity	
						Cochran'sQ	P-value
THBS2	2	IVW	0.211 (0.047-0.939)	-1.554	0.041	0.116	0.733

For exposures with more than two SNPs, IVW analysis indicated that *TMEM200A* was linked to decreased NPC risk (OR = 0.288, 95% CI: 0.106-0.781, P = 0.014), while *VPREB3* was associated with increased risk (OR = 2.903, 95% CI: 1.192-7.068, P = 0.019). (Table 3)

Table 3 Mendelian randomization results for *TMEM200A* and *VPREB3*

Exposure	Snp	Methods	OR(95%CI)	Beta	P-value
TMEM200A	7	MR Egger	0.245(0.066-0.904)	-1.405	0.088
		Weighted median	0.339(0.111-1.036)	-1.081	0.057
		IVW	0.288(0.106-0.781)	-1.244	0.014
		Simple mode	0.646(0.089-4.656)	-0.437	0.679
		Weighted mode	0.358(0.113-1.132)	-1.025	0.131
VPREB3	14	MR Egger	6.517(1.367-31.049)	1.874	0.036
		Weighted median	3.754(1.314-10.720)	1.322	0.013
		IVW	2.903(1.192-7.068)	1.066	0.019
		Simple mode	1.070(0.110-10.355)	0.068	0.954
		Weighted mode	3.646(1.136-11.704)	1.299	0.049

Heterogeneity testing revealed no significant heterogeneity between *TMEM200A* and *VPREB3* in relation to nasopharyngeal carcinoma (Cochran's Q p > 0.05). Furthermore, pleiotropy analysis indicated no detectable pleiotropic effect of these two genes on nasopharyngeal carcinoma (p > 0.05). (Table 4)

Table 4 Results of heterogeneity and pleiotropy tests for *TMEM200A* and *VPREB3*

Exposure	Snp	Methods	Heterogeneity			MR-PRESSO	Horizontal pleiotropy		
			Cochran'sQ	Qdf	P-value	P-value	Egger intercept	Se	P-value
TMEM200A	7	MR Egger	2.893	5	0.716	0.864	0.047	0.127	0.722
		IVW	3.034	6	0.804				
VPREB3	14	MR Egger	15.037	12	0.239	0.239	-0.111	0.091	0.244
		IVW	16.917	13	0.203				

Single-Cell Expression Profiling of Key Genes

To evaluate batch effects and determine an appropriate clustering resolution for downstream single-cell analyses, dimensionality reduction and clustering diagnostics were performed. Prior to batch correction, principal component analysis (PCA) revealed that cells were primarily separated according to dataset and sample origin, indicating the presence of pronounced batch effects (Figure 2A). After integration using Harmony, cells from different datasets and samples were well mixed in the low-dimensional space, demonstrating effective batch correction (Figure 2B).

To further assess clustering behavior across different resolutions, clustering results were systematically examined over a range of resolution parameters using a cluster tree visualization. The clustering tree showed that clusters underwent progressive and stable refinement as resolution increased. At a resolution of 0.2, major cell populations were consistently preserved with limited cluster fragmentation, providing a balance between cluster stability and granularity. Based on this evaluation, a resolution of 0.2 was selected for subsequent clustering and downstream analyses (Figure 2C).

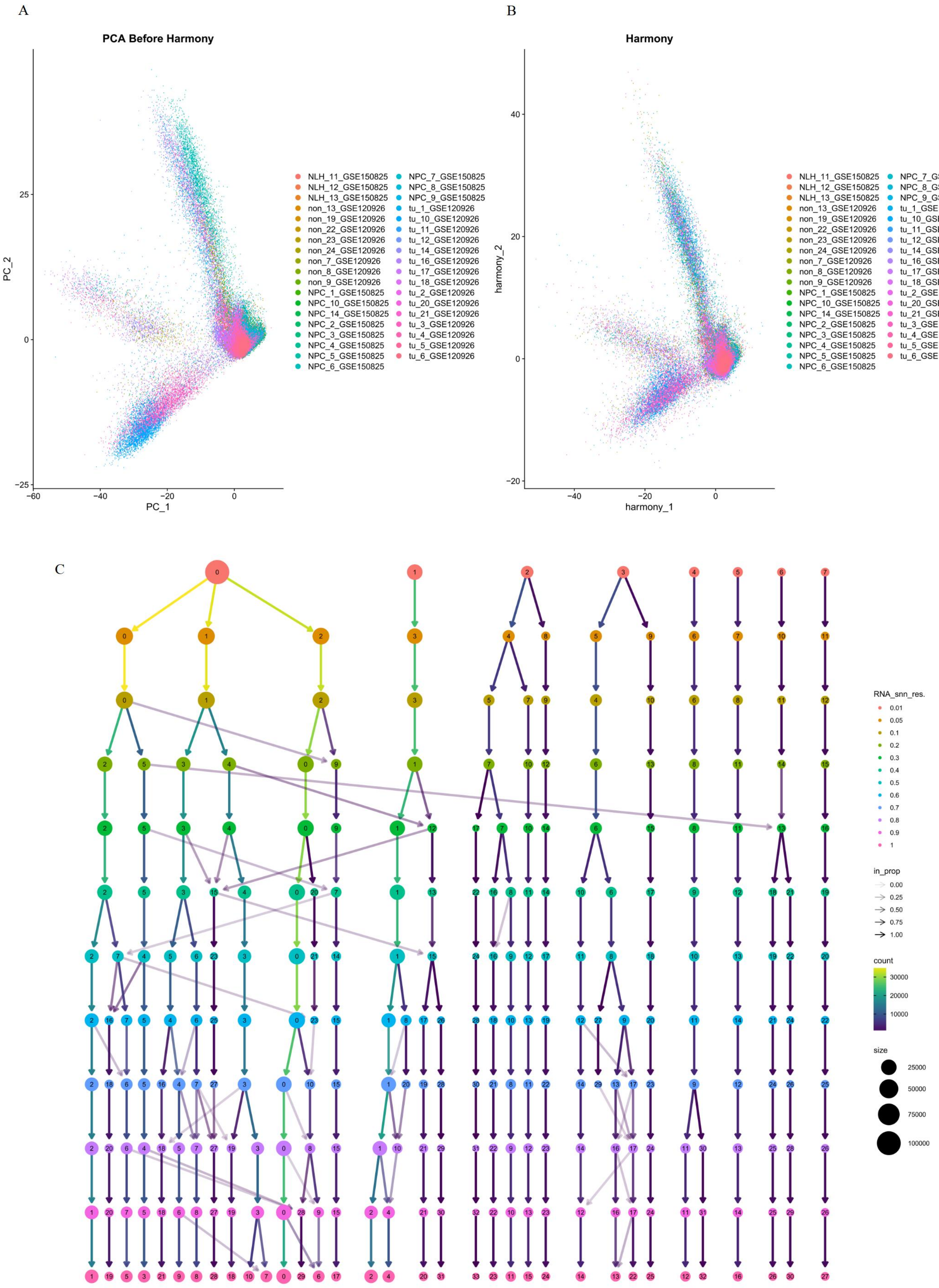

Figure2. (A)Principal component analysis (PCA) of single-cell transcriptomes before batch correction, colored by sample and dataset, showing clear batch-driven separation. (B)Low-dimensional embedding after Harmony integration, demonstrating effective batch correction and improved mixing of cells across datasets and samples. (C)Cluster tree visualization across multiple resolution parameters generated using the clustree approach. Nodes represent clusters at different resolutions, with node size indicating cell numbers and edges showing cluster relationships across resolutions.

After quality control, 157,289 cells from two single-cell RNA-seq datasets were retained for analysis. Unsupervised clustering at a resolution of 0.2 identified 11 distinct cell populations. These clusters were annotated based on established marker genes as follows: B cells (*CD79A*), CD4T cells (*CD3D*, *CD4*), CD8T cells (*CD8A*), Myeloid cells (*LYZ*), plasma B cells (*SDC1*), epithelial cells

(*EPCAM*), fibroblasts (*COL1A1*), endothelial cells (*VWF*), mast cells (*TPSAB1*), plasmacytoid dendritic cells (*LILRA4*), and Cycling cells (*MKI67*).

As shown in Figure 3A and 3C, single-cell RNA sequencing data were annotated into distinct cell populations, and the expression of canonical marker genes confirmed the identity of each annotated cell type. Compared with non-tumor tissues, nasopharyngeal carcinoma tissues exhibited a marked remodeling of immune cell composition, characterized by increased proportions of CD8^{+}T cells, CD4^{+}T cells, myeloid cells, and epithelial cells, along with a relative reduction in B cells (Figure 3B) . In addition, tumor tissues showed enrichment of fibroblast and endothelial cell populations, suggesting expansion of stromal and vascular components within the NPC tumor microenvironment.

To further characterize the cellular context of MR-identified genes, their expression patterns were examined across annotated cell types and between tumor and non-tumor samples. Dot plot analysis revealed distinct cell-type-specific expression profiles of the five candidate genes (Figure 3D). *PSPH* was predominantly expressed in epithelial cells. *TMEM200A* and *THBS2* expression was enriched in fibroblasts; notably, *TMEM200A* was also detected in CD4+ and CD8+ T cells. *MFSD4* was mainly found in endothelial cells, and *VPREB3* was highly expressed in B cells, with a high percentage of cells expressing it. Comparison between tumor and non-tumor samples further demonstrated differential expression of these genes at the cellular level (Figure 3E). In particular, *VPREB3* exhibited higher expression and a larger fraction of expressing cells in non-tumor samples, whereas other genes showed more subtle but cell-type-dependent expression differences, indicating that MR-identified genetic signals are reflected in distinct cellular expression patterns within the tumor microenvironment.

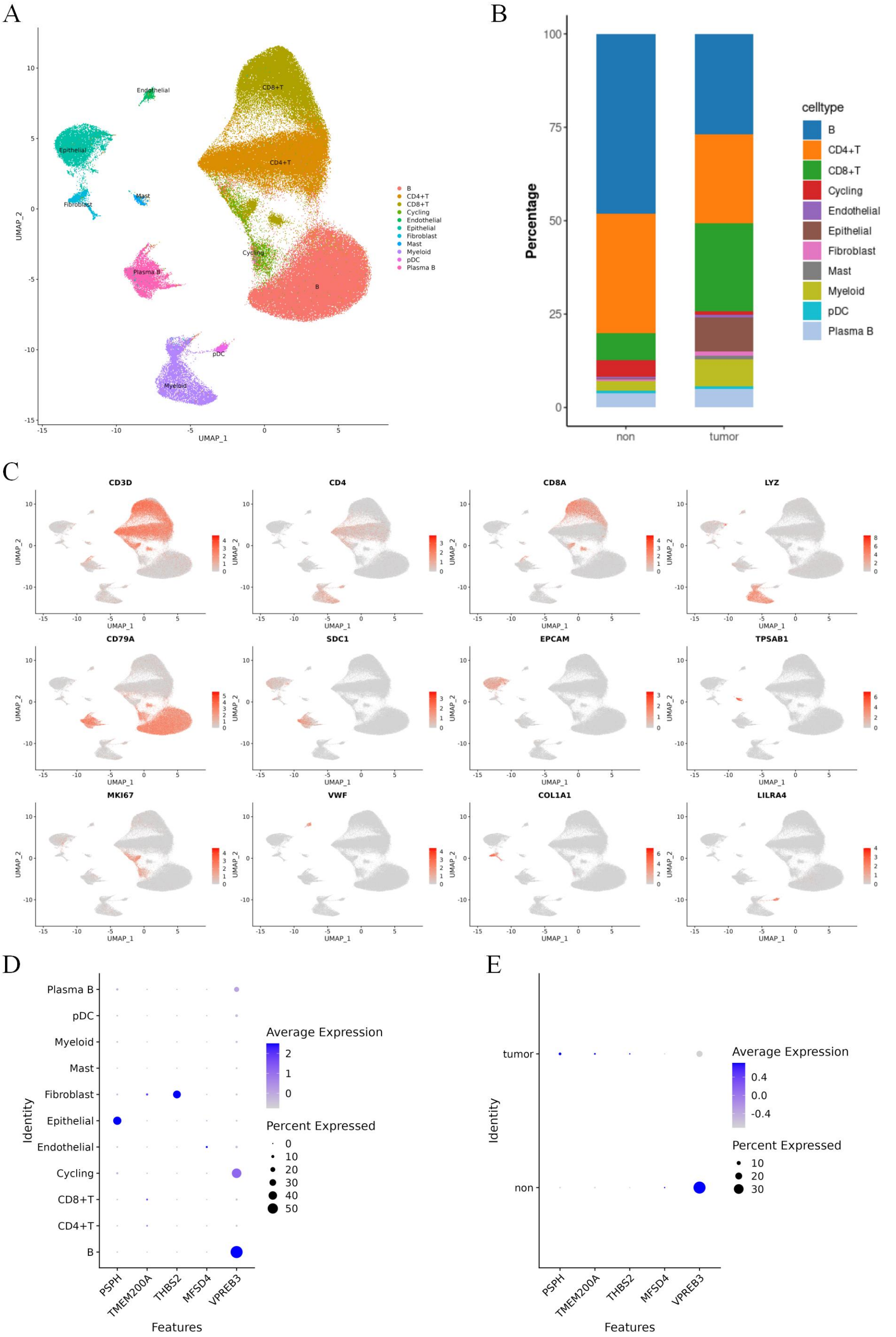

Figure3. (A)UMAP projection of integrated single-cell RNA-seq data showing major annotated cell types based on canonical marker genes. (B)Relative proportions of cell types in non-tumor and tumor samples. (C)Feature plots of representative marker genes used for cell-type annotation. (D)Dot plot showing the expression of *PSPH*, *TMEM200A*, *THBS2*, *MFSD4*, and *VPREB3* across cell types. (E)Dot plot comparing the expression of the five genes between non-tumor and tumor samples.

To further characterize the expression patterns of the Mendelian randomization-identified genes under disease conditions, stratified comparisons were performed between nasopharyngeal carcinoma (NPC) samples and control samples within the primary cell types expressing these genes (Figure 4). In epithelial cells, the risk gene *PSPH* was significantly upregulated in NPC samples compared with controls (Wilcoxon test, $P < 0.0001$), indicating its aberrant activation in tumor-associated epithelial cells. In fibroblasts, the protective gene *THBS2* exhibited significant expression changes in NPC tissues ($P < 0.0001$), and *TMEM200A* also showed a statistically significant difference in this cell population ($P < 0.05$), suggesting pronounced molecular remodeling of stromal cells during NPC development. Among immune cells, *TMEM200A* displayed a modest yet significant expression difference in CD8^{+}T cells ($P < 0.05$), implying its potential involvement in regulating tumor-related immune status. Furthermore, in B-cell lineages, the risk gene *VPREB3* showed significant differential expression between NPC and control samples in B cells, cycling cells, and plasma cells (all $P < 0.0001$), demonstrating consistent disease-associated expression alterations. In contrast, *MFSD4* expression in endothelial cells did not differ significantly between NPC and control samples ($P > 0.05$), despite its clear cell type-specific expression pattern at the single-cell level. These results indicate that the expression imbalance of MR-identified pathogenic and protective genes in NPC is not uniformly driven by all cell types, but is primarily mediated by specific immune and stromal cell populations.

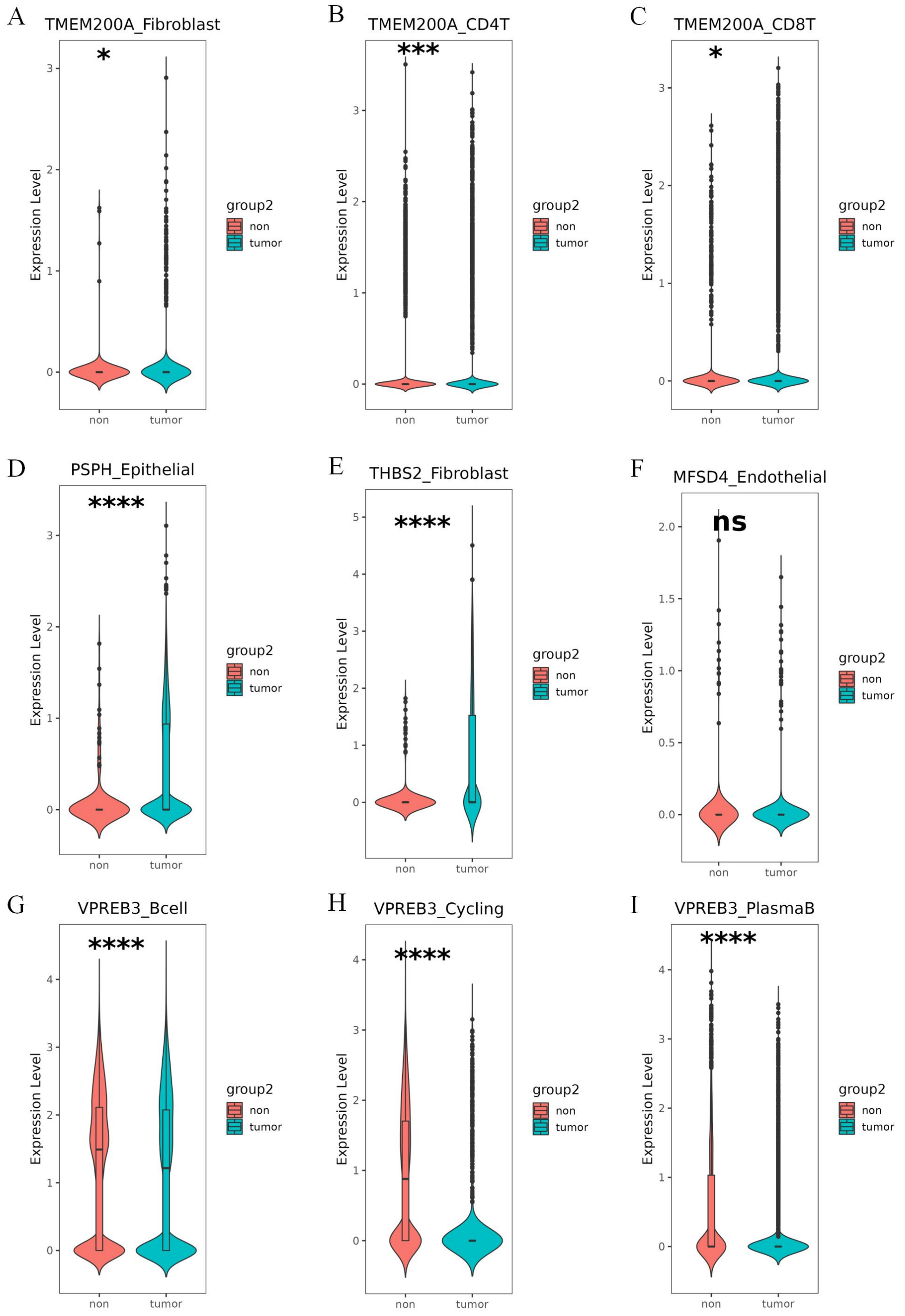

Figure 4. Cell-type-specific differential expression of MR-identified genes between non-tumor and tumor samples. Violin plots showing the expression differences of key genes between non-tumor and tumor groups within specific cell types. **(A–C)** Expression of *TMEM200A* in fibroblasts(A), $CD4^+$ T cells (B), and $CD8^+$ T cells (C). **(D)** Expression of *PSPH* in epithelial cells. **(E)** Expression of *THBS2* in fibroblasts. **(F)** Expression of *MFSD4* in endothelial cells. **(G–I)** Expression of *VPREB3* in B cells (G), cycling cells (H), and plasma B cells (I). For each panel, violin plots represent the distribution of gene expression levels, with overlaid boxplots indicating median and interquartile ranges. Statistical significance between non-tumor and tumor groups was assessed using the Wilcoxon rank-sum test. Significance levels are indicated as $P < 0.05$ (*), $P < 0.01$ (**), $P < 0.001$ (***), $P < 0.0001$ (****), and "ns" denotes not significant.

Discussion

By integrating multivariable Mendelian randomization (MR) analysis with single-cell RNA sequencing data, this study systematically elucidated the pathogenic genes underlying nasopharyngeal carcinoma (NPC) and their potential mechanisms from the perspectives of genetic causal inference and cellular resolution. The MR analysis identified five key genes with causal associations with NPC development, among which *TMEM200A* and *THBS2* acted as protective factors, while *VPREB3*, *MFSD4*, and *PSPH* were identified as risk factors. Subsequent single-cell transcriptomic analysis not only delineated the cell type–specific expression patterns of these genes within the tumor microenvironment but also revealed their expression dysregulation between NPC tissues and control tissues, thereby providing a clear cellular context for the genetic risk signals.

TMEM200A belongs to the transmembrane protein family, and its role in tumorigenesis remains poorly understood. In this study, MR analysis confirmed *TMEM200A* as a protective factor for NPC (OR = 0.288, 95% CI: 0.106–0.781, P = 0.014). Comparative analysis at the single-cell level further showed that *TMEM200A* was primarily expressed in fibroblasts and T cells, with significant differences between NPC and control samples. Fibroblasts are key regulators of extracellular matrix remodeling and immune cell recruitment in the tumor microenvironment[12,13], while T cells play a central role in antitumor immunity and immune surveillance[14]. The relatively high expression of *TMEM200A* in these two cell types suggests that it may suppress the initiation and progression of NPC by modulating fibroblast–T cell interactions and promoting an immune-supportive microenvironment [15].

THBS2 encodes an extracellular matrix–associated protein involved in cell–cell and cell–matrix interactions, and plays an important role in regulating angiogenesis[16]. This study identified *THBS2* as a significant protective factor for NPC at the genetic level (OR = 0.211, 95% CI: 0.047–0.939, P = 0.041). Single-cell analysis showed that *THBS2* was predominantly expressed by fibroblasts and was significantly dysregulated in NPC compared with control tissues. Previous studies have shown that *THBS2* acts as an endogenous angiogenesis inhibitor that restricts tumor neovascularization by antagonizing pro- angiogenic signals[17]. Combined with our findings, we propose that *THBS2* may attenuate microenvironmental support for tumor growth and metastasis by inhibiting fibroblast-mediated angiogenesis and stromal remodeling.

MFSD4 was identified as a high-risk gene for NPC (OR = 11.020, 95% CI: 1.637–74.150, P = 0.014) and was primarily expressed in endothelial cells at the single-cell level. Tumor-associated endothelial cells play a crucial role in angiogenesis, tumor invasion, and metastasis[16,18]. Notably, endothelial dysfunction is often associated with sustained activation of pro-angiogenic signaling pathways such as *VEGF*[19]. Although the precise molecular function of *MFSD4* remains incompletely understood, its high expression in endothelial cells suggests that this gene may promote NPC angiogenesis by enhancing endothelial responsiveness to pro-angiogenic signals, thereby indirectly facilitating *VEGF* pathway activation. This hypothesis offers a plausible explanation for the role of *MFSD4* as a genetic risk factor and suggests that the MFSD4–VEGF axis may represent a potential target for future functional studies and therapeutic intervention.

PSPH was significantly upregulated in epithelial cells, providing direct cytological evidence for its role as a risk gene (OR = 3.942, 95% CI: 1.135–13.685, P = 0.031) in nasopharyngeal carcinoma (NPC). As a key enzyme in the serine biosynthesis pathway[20], aberrant activation of *PSPH* may enhance the anabolic capacity of tumor cells and support their proliferation by promoting the supply of nucleotides and amino acids. Furthermore, previous studies have shown that PSPH can enhance tumor invasiveness through signaling pathways such as *MAPK*[21]. Its specific elevation in NPC epithelial cells suggests that metabolic reprogramming may be an important mechanism underlying its oncogenic effects.

VPREB3 is involved in early B-cell development and was identified as a risk gene for NPC (OR = 2.903, 95% CI: 1.192–7.068, P = 0.019). Single-cell analysis revealed that *VPREB3* was mainly expressed in B cells and plasma cells and displayed aberrant expression in NPC tissues. Previous research suggests that B cells can promote tumorigenesis by forming circulating immune complexes[22] and activating Fc receptors on myeloid cells, thereby inducing chronic inflammation[23]. Aberrant expression of *VPREB3* may enhance pro-tumor inflammatory microenvironments or immune evasion by interfering with B-cell receptor signaling or affecting B-cell differentiation states.

In summary, our findings demonstrate that the genetic susceptibility to nasopharyngeal carcinoma is primarily mediated through functional imbalances across distinct cell types within the tumor microenvironment, rather than being driven solely by intrinsic alterations in tumor cells. Integrating Mendelian randomization with single-cell transcriptomic data enables precise mapping of genetic causal signals to specific cellular populations, thereby providing a more refined cellular biological framework for understanding the pathogenesis of nasopharyngeal carcinoma.

Limitations

Several limitations should be considered in interpreting our findings. MR estimates may be biased in the presence of horizontal pleiotropy, although sensitivity analyses in this study did not detect significant pleiotropic effects. Furthermore, while MR provides evidence for causal inference, the conclusions require validation through functional experiments and clinical studies. The GWAS and eQTL data used in this study were primarily derived from European populations, which may limit the generalizability of the results to other ethnic groups. Additionally, static eQTL data do not capture dynamic gene regulation over time or context, which may affect the interpretation of gene – disease causality. Future studies leveraging larger eQTL reference panels or tissue-specific regulatory datasets may help refine the causal effect estimate of *MFSD4*. And more diverse cohorts and time-resolved functional genomics data will help refine these findings.

Conclusion

By integrating Mendelian randomization with single-cell RNA sequencing, we identified five genes with causal relevance to nasopharyngeal carcinoma risk and resolved their cell-type-specific expression patterns within the tumor microenvironment. Protective genes were primarily enriched in stromal and immune cells, whereas risk genes localized to B cells, endothelial cells, and epithelial cells, suggesting distinct cellular routes through which genetic susceptibility influences NPC development. This integrative framework bridges genetic causality and disease-relevant cellular biology.

Reference

1. Chen, Y. P., Chan, A. T. C., Le, Q. T., Blanchard, P., Sun, Y., & Ma, J. (2019). Nasopharyngeal carcinoma. Lancet (London, England), 394(10192), 64–80.
2. Smith, G. D., & Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?. International journal of epidemiology, 32(1), 1–22.
3. Bao, Q., Zou, J., Wang, C., & Wang, H. (2025). Mendelian randomization analysis reveals potential association between allergic rhinitis and nasopharyngeal carcinoma. Discover oncology, 16(1), 799.
4. Yi, T., & Lin, S. (2024). The protective role of vitamin d in nasopharyngeal carcinoma: insights from Mendelian randomization and meta-analysis. Discover oncology, 15(1), 637.
5. Li, W., Zhang, Z., Xie, B., He, Y., He, K., Qiu, H., Lu, Z., Jiang, C., Pan, X., He, Y., Hu, W., Liu, W., Que, T., & Hu, Y. (2024). HiOmics: A cloud-based one-stop platform for the comprehensive analysis of large-scale omics data. Computational and structural biotechnology journal, 23, 659–668.
6. Burgess, S., Thompson, S. G., & CRP CHD Genetics Collaboration (2011). Avoiding bias from weak instruments in Mendelian randomization studies. International journal of epidemiology, 40(3), 755–764.
7. Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. International journal of epidemiology, 44(2), 512–525.
8. Bowden, J., Davey Smith, G., Haycock, P. C., & Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genetic epidemiology, 40(4), 304–314.
9. Hartwig, F. P., Davey Smith, G., & Bowden, J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. International journal of epidemiology, 46(6), 1985–1998.
10. Greco M, F. D., Minelli, C., Sheehan, N. A., & Thompson, J. R. (2015). Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. Statistics in medicine, 34(21), 2926–2940.
11. Burgess, S., & Thompson, S. G. (2017). Interpreting findings from Mendelian randomization using the MR-Egger method. European journal of epidemiology, 32(5), 377–389.
12. Sahai, E., Astsaturov, I., Cukierman, E., DeNardo, D. G., Egeblad, M., Evans, R. M., Fearon, D., Greten, F. R., Hingorani, S. R., Hunter, T., Hynes, R. O., Jain, R. K., Janowitz, T., Jorgensen, C., Kimmelman, A. C., Kolonin, M. G., Maki, R. G., Powers, R. S., Puré, E., Ramirez, D. C., … Werb, Z. (2020). A framework for advancing our understanding of cancer-associated fibroblasts. Nature reviews. Cancer, 20(3), 174–186.
13. Zhang D., Qin L., Liu Q., et al. JNK1 downregulation confers fibroblasts with a mammary epithelial cell fate. iCell, 2025 Oct.24; 2(2).https://dx.doi.org/10.71373/PMKJ7669
14. Yang, W., Liu, S., Mao, M., Gong, Y., Li, X., Lei, T., Liu, C., Wu, S., & Hu, Q. (2024). T-cell infiltration and its regulatory mechanisms in cancers: insights at single-cell resolution. Journal of experimental & clinical cancer research : CR, 43(1), 38.
15. Liao, X., Wang, W., Yu, B., & Tan, S. (2022). Thrombospondin-2 acts as a bridge between tumor extracellular matrix and immune infiltration in pancreatic and stomach adenocarcinomas: an integrative pan-cancer analysis. Cancer cell international, 22(1), 213.
16. Cutler, A. A., Pawlikowski, B., Wheeler, J. R., Dalla Betta, N., Elston, T., O'Rourke, R., Jones, K., & Olwin, B. B. (2022). The regenerating skeletal muscle niche drives satellite cell return to quiescence. iScience, 25(6), 104444.
17. Sun, R., Wu, J., Chen, Y., Lu, M., Zhang, S., Lu, D., & Li, Y. (2014). Down regulation of Thrombospondin2 predicts poor prognosis in patients with gastric cancer. Molecular cancer, 13, 225.
18. Fang, J., Lu, Y., Zheng, J., Jiang, X., Shen, H., Shang, X., Lu, Y., & Fu, P. (2023). Exploring the crosstalk between endothelial cells, immune cells, and immune checkpoints in the tumor microenvironment: new insights and therapeutic implications. Cell death & disease, 14(9), 586.
19. Shi, Y., Lei, K., Jia, Y., Ni, B., He, Z., Bi, M., Wang, X., Shi, J., Zhou, M., Sun, Q., Wang, G., Chen, D., Shu, Y., Liu, L., Guo, Z., Liu, Y., Yang, J., Wang, K., Xiao, K., Wu, L., … Wu, C. (2021). Bevacizumab biosimilar LY01008 compared with bevacizumab (Avastin) as first-line treatment for Chinese patients with unresectable, metastatic, or recurrent non-squamous non-small-cell lung cancer: A multicenter, randomized, double-blinded, phase III trial. Cancer communications (London, England), 41(9), 889–903.
20. Liao, L., Ge, M., Zhan, Q., Huang, R., Ji, X., Liang, X., & Zhou, X. (2019). PSPH Mediates the Metastasis and Proliferation of Non-small Cell Lung Cancer through MAPK Signaling Pathways. International journal of biological sciences, 15(1), 183–194.
21. Yao, M., Xie, Y., Huang, M., Han, X., Zhou, Y., Tao, M., Liu, C., Zhao, Y., Zhang, C., & Gao, Y. (2025). PSPH promotes the proliferation and metastasis of esophageal squamous cell carcinoma through MAPK signaling pathways. American journal of cancer research, 15(4), 1919–1931.
22. Zhong, F., Chen, J., Lu, T., Zhang, L., Liu, Z., Guan, C., Xiong, X., Gong, X., & Li, J. (2025). Infiltrating B-cell subtypes and associated hub genes in nasopharyngeal carcinoma identified from integrated single-cell, bulk RNA-sequencing, and immunohistochemical data. Hereditas, 162(1), 48.
23. Largeot, A., Pagano, G., Gonder, S., Moussay, E., & Paggetti, J. (2019). The B-side of Cancer Immunity: The Underrated Tune. Cells, 8(5), 449.

Author Contributions

YuxiaoHe, BinHan and YanlingHu designed the experiment and drafted and revised the manuscript. JianpingZhou and XuanyuPan, Qingniao Zhou, TingtingLiu retrieved data from databases. TingweiLv, XingraoLi, YongjianZhou, and YaniYang performed data analysis. All authors contributed to the manuscript and approved the submitted version.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Department of Science and Technology of Guangxi Zhuang Autonomous Region (Guike 2023AB04032), the China National Center for Biotechnology Development (2023YFC05400), the National Natural Science Foundation of China (No. 32260164) and Natural Science Foundation of Guangxi Zhuang Autonomous Region (No. 2023GXNSFAA026080).

Ethics statement

This study analyzed existing public data from GEO, OpenGWAS, and FinnGen databases. All original studies provided ethical approval. No additional ethics approval was required for this secondary analysis.

Consent Comments

Not applicable. This study is a secondary analysis of existing, de-identified public data. All original studies that contributed data to the used public databases obtained informed consent from participants.

Data availability statement

The original datasets presented in this study are openly available in public repositories. The bulk RNA-seq data (GSE53819, GSE12452) and single-cell RNA-seq data (GSE150825, GSE120926) can be found in the Gene Expression Omnibus (GEO) database at https://www.ncbi.nlm.nih.gov/geo/. The summary-level GWAS data for eQTLs and nasopharyngeal carcinoma were sourced from the OpenGWAS database http://www.ieu.uk/data/ and the FinnGen consortium https://www.finngen.fi/en/access_results, respectively.