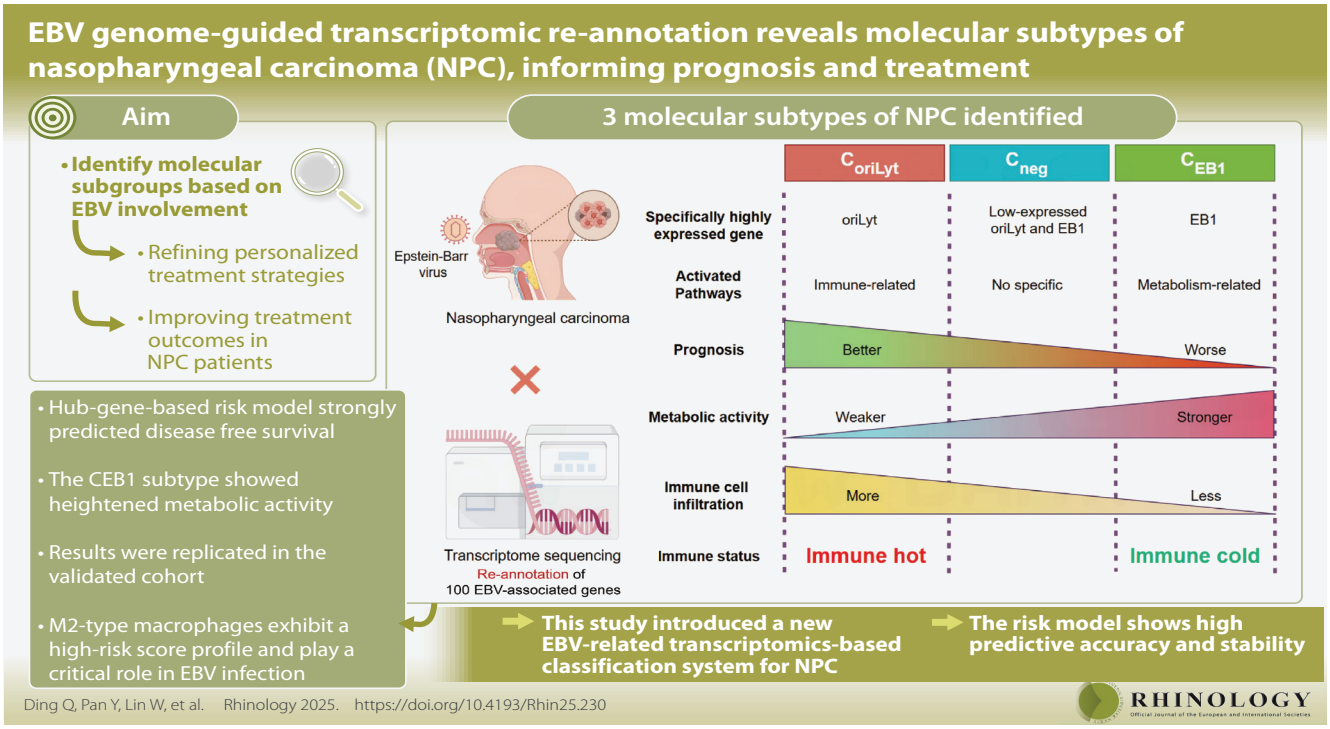
EBV genome-guided transcriptomic re-annotation reveals molecular subtypes of nasopharyngeal carcinoma informing prognosis and treatment

Qin Ding[1,2#], Yuhui Pan[1,2#], Wanzun Lin[1,2], Hanxuan Yang[1,2], Xin Chen[1,2], Haolan Li[1,2], Youliang Weng[1,2*], Sufang Qiu[1,2*]

Rhinology 64: 1, 0 - 0, 2026

https://doi.org/10.4193/Rhin25.274



Abstract

Background: Non-keratinizing nasopharyngeal carcinoma (NPC) is closely related to Epstein-Barr virus (EBV) infection. Patients with NPC often exhibit diverse treatment responses due to tumor heterogeneity. Thus, identifying molecular subgroups based on EBV involvement holds promise for refining personalized treatment strategies and improving treatment outcomes in NPC patients. **Methods**: 193 treatment-naïve NPC specimens with comprehensive clinical and pathological data were procured from Fujian Cancer Hospital. RNA sequencing was employed to acquire the gene expression profiles, followed by the re-annotation of 100 EBV-associated genes leveraging the EBV sequence. Molecular subtypes were conducted via consensus clustering, with an external NPC cohort serving as a validation dataset. Scissor method was applied to identify survival-associated cell subpopulations from single-cell data, following comprehensive bioinformatic analyses. **Results**: Three molecular subtypes of NPC—C_{oriLyt}, C_{neg}, and C_{EB1}—were identified, each with specific clinical profiles. The CEB1 subtype is distinguished by its heightened metabolic activity and immunosuppressive environment. A hub-gene-based risk model for these subtypes strongly predicted disease-free survival, with replicated results in the validated cohort. The model's predictive accuracy was high, with areas under the curve for 1, 3, and 5-year survival rates at 0.79, 0.86, and 0.88, respectively. M2-type macrophages exhibit a high-risk score profile and play a critical role in EBV infection, with prominent activation of the TNF-II and TGF-β signaling pathways. **Conclusions**: This study introduced a new EBV-related transcriptomics-based classification system for NPC that showed great promise in predicting patient survival outcomes.

Key words: nasopharyngeal carcinoma, Epstein-Barr virus, transcriptome sequencing, molecular subtype, risk model, immunotherapy

Introduction

Nasopharyngeal carcinoma (NPC) is common in Southeast Asia, particularly southern China, with metastasis and recurrence being leading causes of death [1–3]. The anatomically-based tumor-node-metastasis (TNM) staging system is widely used but inadequate due to tumor heterogeneity, leading to uncertain patient outcomes [4–6]. Thus, precise molecular subtypes are essential for predicting clinical outcomes and informing therapeutic management, including risk-adapted treatment intensity and selection of targeted or immunotherapeutic approaches. Epstein-Barr virus (EBV) is strongly associated with most non-keratinizing subtypes of NPC and the degree of EBV involvement varies across histological and molecular subtypes [7]. Serum levels of EBV-specific immunoglobulin A (IgA) antibodies—targeting the viral capsid antigen (VCA) and early antigen (EA)—are significantly elevated in NPC patients compared to healthy individuals [8]. Consequently, considering the molecular typing of NPC from the perspective of EBV infection is a feasible concept based on etiology and clinical manifestations. While the WHO subclassification system is commonly used for NPC, more clinicians recognize its limitations in predicting chemotherapy and radiotherapy efficacy [9,10]. Next-generation sequencing tools have enabled the creation of large-scale data profiles in numerous malignant neoplasms, enhancing systematic and accurate tumor characterization [11,12]. Considering the significant impact of molecular events on patient prognosis and treatment regimens, as well as the prominent role of EBV from both etiological and clinical perspectives, it is crucial to identify and characterize molecular subtypes based on the expression profiles of individual tumors with EBV sequence re-annotation. NPC can be classified into three molecular subgroups by microRNA (miRNA) expression, but their distinct pathway enrichments are yet to be fully understood [13]. In addition, an epigenomic mapping study revealed global methylation changes within subtypes, but with limited sample size [14]. A recent NPC classification suggested three molecular subtypes (immune, proliferative, and metabolic), excluding EBV [15].

The current understanding of NPC heterogeneity lacks insights into the high-risk factor EBV. And there is a need to translate transcriptomic findings into improved treatment approaches for NPC. Our study endeavors to analyze gene expression patterns in NPC patients based on EBV sequences, identify novel molecular subtypes for NPC classification, and assess their clinicopathological characteristics.

Materials and methods

Clinical sample collection

Fresh tumor tissues were prospectively collected at the time of diagnostic nasopharyngeal biopsy from NPC patients treated at Fujian Cancer Hospital between January 2015 and January 2018. Immediately after acquisition, specimens were cryopreserved in liquid nitrogen for long-term storage. This study represents a retrospective transcriptomic analysis of these prospectively collected biospecimens. Patients were TNM-staged, and their clinical characteristics are summarized in Table 1 and Table S1. Ethics approval (No. K2022-084-01) and informed consent were obtained from each participant. External validation cohort GSE102349 was retrieved from the Gene Expression Omnibus (GEO) database.

Quantification of plasma EBV DNA

The experimental procedures of this part are detailed in the supplementary materials.

Transcriptome sequencing

The experimental procedures of this part are detailed in the supplementary materials.

Transcriptome analysis based on EBV reference genome sequences

Low-quality reads were filtered out of the sequencing raw data using fastp [16]. The main filtering criteria included 1) filtering reads that did not contain splice sequences or contained N, 2) cutting splice sequences, 3) filtering fragments with an average base mass of less than 20 in a window of 5bp, and 4) filtering reads with a final length of less than 50bp. downloaded from NCBI EBV (GCF_002402265.1) reference genome sequence, use HISAT [17] to create an index file of the reference genome, and compare the high-quality sequencing data to this genome, and the final result contains a BAM file that uniquely compares to the genome. Gene expression was calculated using HTSeq-count tool [18] to obtain the sequencing counts for each gene in each sample, and the corresponding general transfer format (GTF) files were obtained from https: //ebv.wistar.upenn.edu/downloadstatic/ebv.custom.gtf. The expression of each gene was obtained by FPKM and TPM normalisation.

Single-cell (scRNA)-seq data source and processing

The experimental procedures of this part are detailed in the supplementary materials.

Identification of survival-associated single cells using the scissor algorithm

The experimental procedures of this part are detailed in the supplementary materials.

Risk score assessment and cellular communication analysis in single-cell data

The experimental procedures of this part are detailed in the supplementary materials.

Detecting differentially expressed genes (DEGs) between

Table 1. Clinical features profile of NPC patients.

Characteristics	Male	Female	p-value
n	136	57	
Age, mean ± SD	48.824 ± 10.827	47.421 ± 9.8488	0.401
Pathological type, n (%)			0.528
Non-keratinizing undifferentiated	133 (68.9%)	57 (29.5%)	
Keratinizing moderately differentiated	1 (0.5%)	0 (0%)	
Keratinizing poorly differentiated	2 (1%)	0 (0%)	
T, n (%)			0.365
1	24 (12.4%)	16 (8.3%)	
2	31 (16.1%)	12 (6.2%)	
3	44 (22.8%)	18 (9.3%)	
4	37 (19.2%)	11 (5.7%)	
N, n (%)			0.403
0	13 (6.7%)	2 (1%)	
1	45 (23.3%)	22 (11.4%)	
2	53 (27.5%)	25 (13%)	
3	25 (13%)	8 (4.1%)	
M, n (%)			1.000
0	128 (66.3%)	54 (28%)	
1	8 (4.1%)	3 (1.6%)	
Stage, n (%)			0.243
I	4 (2.1%)	0 (0%)	
II	24 (12.4%)	15 (7.8%)	
III	50 (25.9%)	23 (11.9%)	
IV	58 (30.1%)	19 (9.8%)	
Induction chemotherapy cycle, n (%)			0.673
0	22 (11.4%)	11 (5.7%)	
1	2 (1%)	2 (1%)	
2	53 (27.5%)	18 (9.3%)	
3	42 (21.8%)	15 (7.8%)	
4	8 (4.1%)	5 (2.6%)	
5	1 (0.5%)	0 (0%)	
6	8 (4.1%)	6 (3.1%)	
EB-DNA before treatment, median (IQR)	704.5 (500, 5922.5)	886 (500, 4840)	0.993
Induction chemotherapy, n (%)			0.599
Yes	114 (59.1%)	46 (23.8%)	
No	22 (11.4%)	11 (5.7%)	
Whether RT is complete			
Yes	136 (70.5%)	57 (29.5%)	
No	0 (0%)	0 (0%)	
Targeted RT, n (%)			0.600
Yes	31 (16.1%)	15 (7.8%)	
No	105 (54.4%)	42 (21.8%)	

continues on next page

Characteristics	Male	Female	p-value
CCRT cycle, n (%)			0.776
0	29 (15%)	15 (7.8%)	
1	18 (9.3%)	9 (4.7%)	
2	76 (39.4%)	29 (15%)	
3	13 (6.7%)	4 (2.1%)	

RT: radiotherapy; Targeted RT: Whether targeted therapy during radiotherapy; IQR: Interquartile Range; CCRT cycle: Cycle of concurrent chemoradiotherapy.

NPC and normal tissues

The experimental procedures of this part are detailed in the supplementary materials.

Gene set variation analysis

The experimental procedures of this part are detailed in the supplementary materials.

Gene Ontology (GO) enrichment analysis

The experimental procedures of this part are detailed in the supplementary materials.

Consensus clustering

The experimental procedures of this part are detailed in the supplementary materials.

Principal component analysis (PCA)

The experimental procedures of this part are detailed in the supplementary materials.

Survival analysis

The experimental procedures of this part are detailed in the supplementary materials.

Weighted gene co-expression network analysis (WGCNA)

WGCNA was executed in R (v4.3.2) using default parameters to construct gene co-expression networks [19–21]. Pearson correlations between modules and phenotypic traits for each subtype were computed and adjusted for FDR using Benjamini-Hochberg. Hub genes, featuring high connectivity within their modules, were pinpointed as crucial regulatory elements.

Evaluation of immune cell infiltration level

The experimental procedures of this part are detailed in the supplementary materials.

Constructing and validating the prognostic risk signature

The experimental procedures of this part are detailed in the supplementary materials.

Chemotherapy and radiotherapy sensitivity evaluation

The experimental procedures of this part are detailed in the supplementary materials.

Immunotherapy response prediction

The experimental procedures of this part are detailed in the supplementary materials.

Results

Consensus clustering identified three subtypes

Transcriptomic exploration was conducted on 193 untreated primary NPC cases from Fujian Cancer Hospital, with clinicopathological characteristics summarized in Table 1. Through EBV genomic re-annotation of existing transcriptomic data, we identified 13 DEGs associated with EBV sequence elements (Figure 1A). These EBV-guided DEGs were then used for consensus clustering via the k-means algorithm, revealing three molecular subtypes of NPC with distinct expression profiles (Figure 1B). Particularly, cluster C1 exhibited pronounced expression of oriLyt, whereas clusters C2 exhibited diminished oriLyt expression but heightened EB1 expression. In contrast, cluster C3 demonstrated minimal expression of both oriLyt and EB1 (oriLyt and EB1 negative) but exhibited expression of other DEGs. Therefore, cluster C1 corresponds to the C_{oriLyt} subtype, C2 to the C_{EB1} subtype, and C3 to the C_{neg} subtype, as shown in Figure 1C. The combined results of PCA analyses revealed a substantial trend of separation among samples from the three clusters, reflecting notable differences and heterogeneity between these subtypes at the transcriptome level and heterogeneity (Figure 1D). We compared EB DNA copy profiles in peripheral blood of clinical patients from these three clusters. The analyses showed that C_{oriLyt} had a higher percentage of elevated EB DNA levels (Figure 1E). Cluster C_{EB1} particularly exhibited increased activity in metabolic pathways, including sphingolipid, arachidonic acid, linoleic acid, and glycerolipid metabolism (Figure 1F). Regarding C_{oriLyt}, it exhibits significant enrichment in immune-related biological processes, including regulation of CD8 positive αβ T cell differentiation, CD40 signaling pathway, activation induced cell death of T cells, and regulation of natural killer cell chemotaxis (Figure 1G).

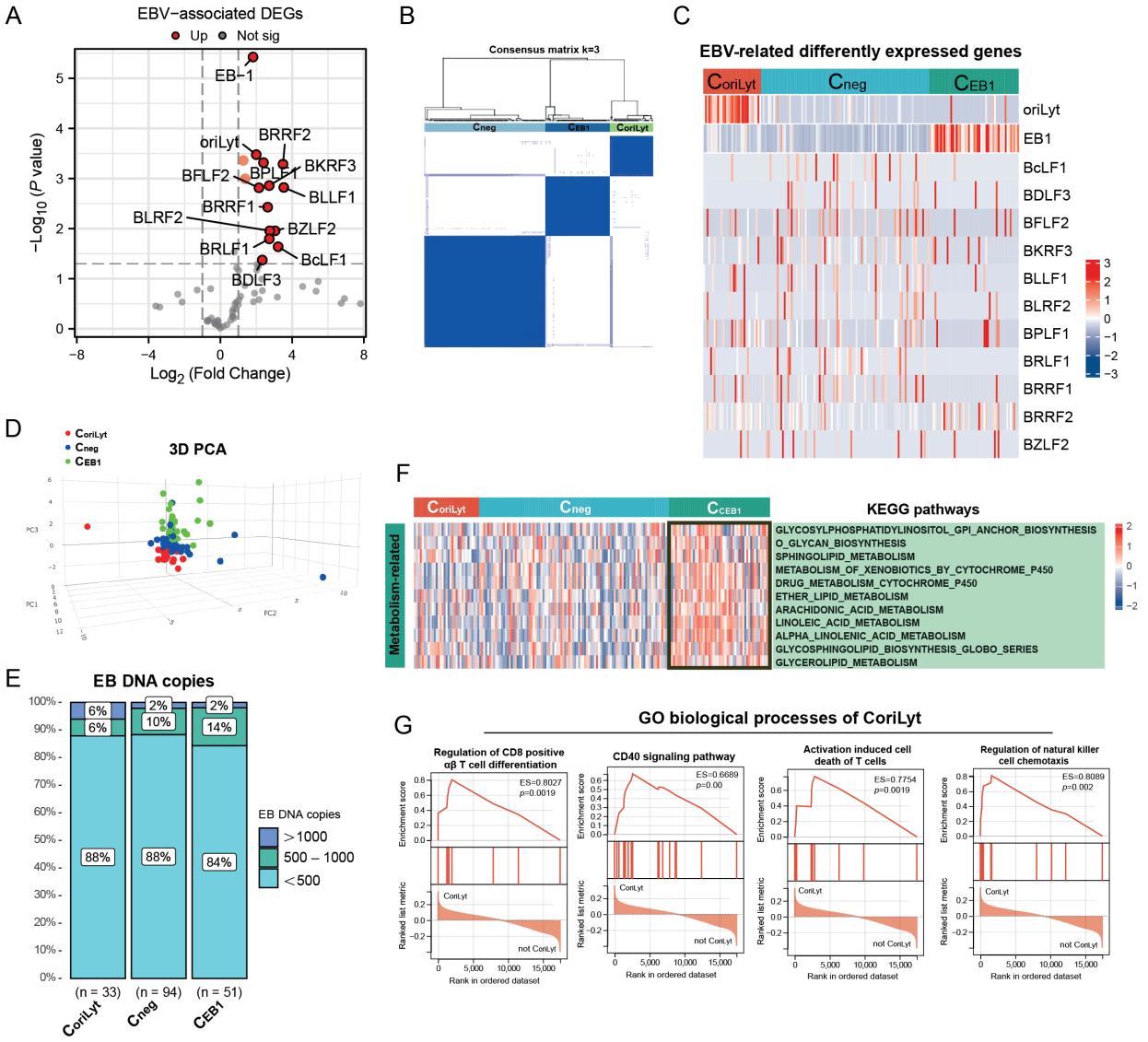


Figure 1. Consensus clustering identifies three molecular subtypes based on EBV sequences in NPC patients. (A) A total of 13 differently expressed genes (DEGs) were identified between tumour tissues and normal tissues of NPC patients based on EBV sequence annotation (n = 193); (B) Heatmap of the consensus clustering scheme (k = 3) in 193 nasopharyngeal carcinoma samples; (C) expression of DEG in the three subtypes; (D) Principle Component Analysis map revealing the different expression in the three subtypes patterns; red dots represent C_{oriLyt} subtype, blue dots represent C_{neg} subtype, and green dots represent C_{EB1} subtype; (E) EB DNA copies of NPC patients in the three subtypes; (F) Heatmap of KEGG pathway scores of C_{oriLyt}, C_{neg}, and C_{EB1} subtypes (n = 193); (G) GO enrichment analysis demonstrating the immunological related pathways of C_{oriLyt} subtype.

Each identified subtype showed distinct clinical feature in patients with NPC

We then investigated whether the three identified subtypes corresponded to distinct clinical characteristics. Maximum standardized uptake value (SUV) of tumor primary lesion (SUV-Tmax) and lymph node (SUV-Nmax) were analyzed from patients undergoing the positron emission tomography–computed tomography (PET-CT) scans. Although SUV values are not definitive indicators of disease severity or prognosis, they reflect underlying metabolic activity and may offer supportive evidence of tumor aggressiveness. In our cohort, both SUV-Tmax and SUV-Nmax showed an increasing trend across the subtypes, with the C_{EB1} subtype exhibiting the highest values, suggesting a potential link with heightened metabolic reprogramming (Figures 2A–B). Notably, patients classified within the C_{EB1} subtype presented with advanced clinical stages, higher recurrence rates, and poor long-term outcomes. In stark contrast, the C_{oriLyt} subtype was characterized by early clinical stages, lower recurrence rates, and more favorable prognosis (Figures 2C–E). Furthermore, stage III-IV patients also revealed a consistent prognostic trend (Figure 2F). This result is demonstrated in the validation dataset (Figure 2G). Moreover, patients with the C_{EB1} subtype

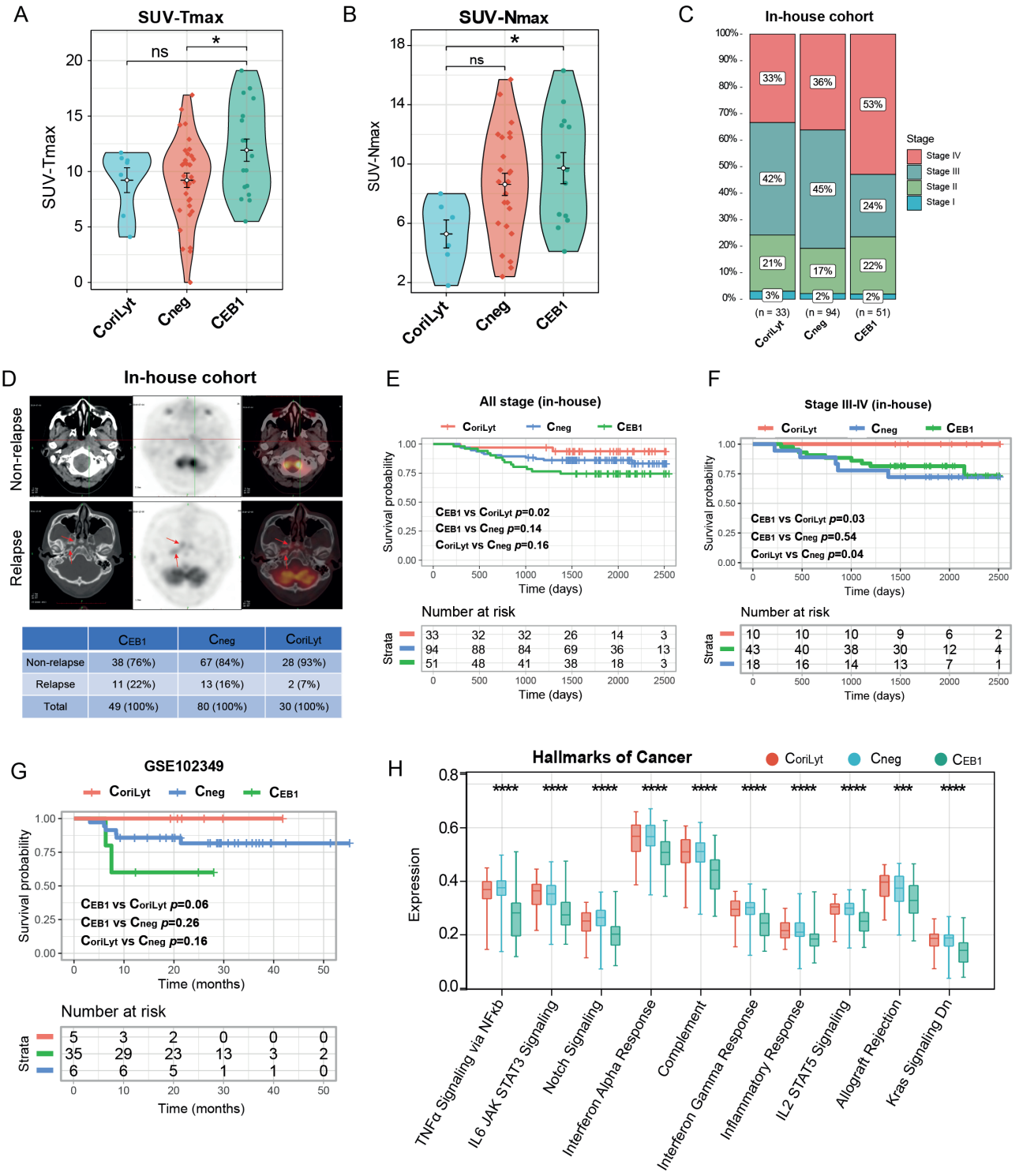


Figure 2. Novel NPC classifications constructed on the basis of EBV sequences show different clinical prognostic features. (A) PET/CT (positron emission tomography/computed tomography) parameter SUV value of the primary lesion of the tumor (SUV-Tmax) in patients with NPC in different subtypes; (B) PET- CT parameter SUV value of the tumor invading the lymph nodes (SUV-Nmax) in patients with NPC in different subtypes; (C) Bar graphs showing the frequency of different subtypes in these subtypes; (D) Relapse rate among these subtypes. NPC relapse was presented by PET/CT; (E – G) Kaplan-Meier disease-free survival curves for all (E) and stage III-IV (F) patients with NPC in the internal cohort and NPC-GSE102349 cohort (G) belonging to C_{oriLyt}, C_{neg}, and C_{EB1} subtypes; (H) Difference of hallmarks of cancer in identified three clusters. ns, $p > 0.05$; *, $p < 0.05$, ***, $p < 0.001$, ****, $p < 0.0001$.

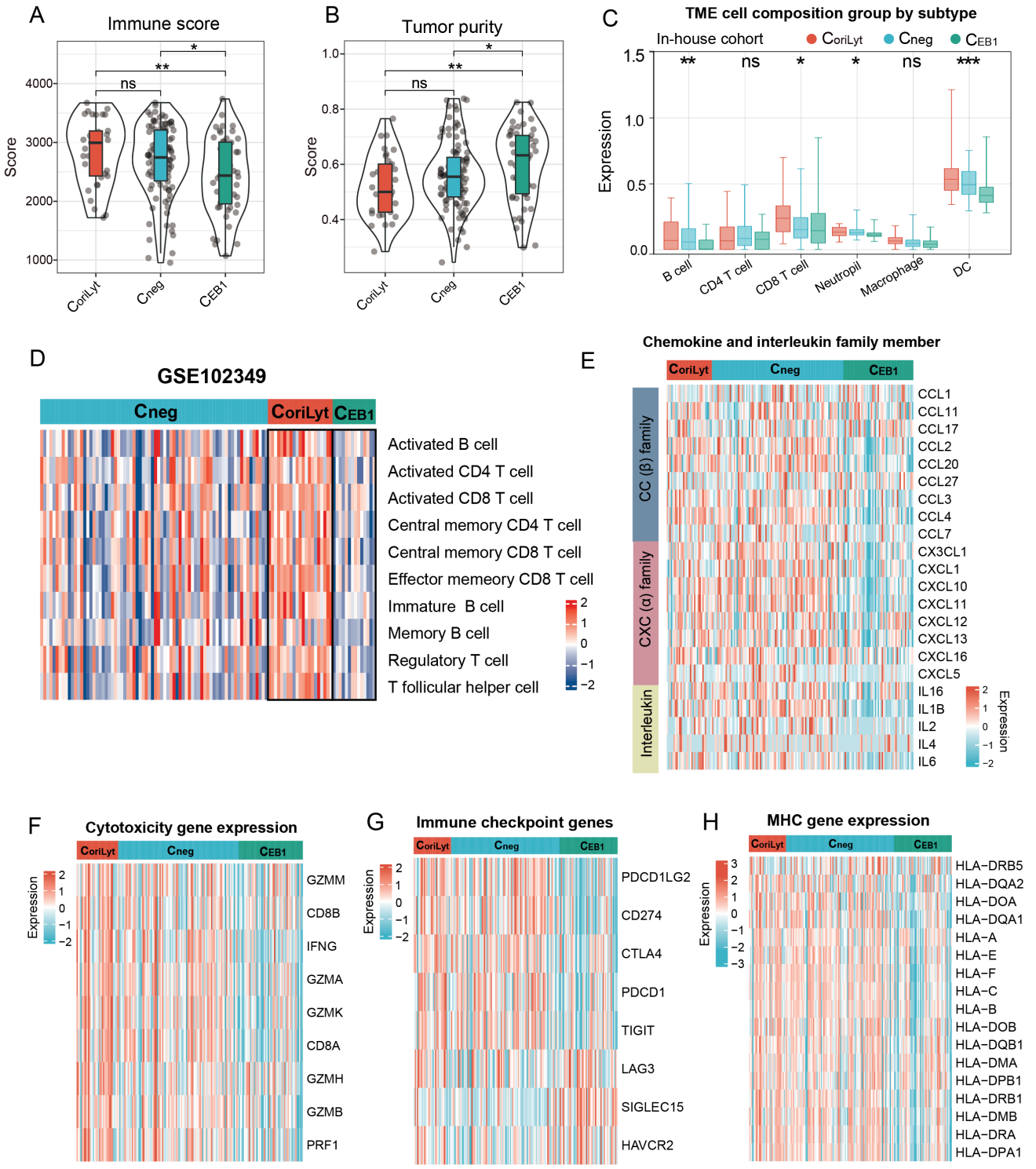


Figure 3. Three subtypes are associated with distinct tumor microenvironments. (A–B) Violin plots showing the median, quartile, and kernel density estimations for each immune score (A) and tumor purity score (B); (C) Box plot of 6 immune cell population score among three subtypes. Red boxes represent C_{oriLyt} subtype, blue boxes represent C_{neg} subtype, and green boxes represent C_{EB1} subtype; (D) Heatmap of immune cell population scores among three subtypes in validation dataset; (E) Heatmap demonstrating the expression of chemokines and interleukin family members among three clusters; (F–H) Heatmaps presenting the differential cytotoxicity gene expression (F), immune checkpoint gene expression (G), and MHC gene expression (H) among the three subtypes. ns, $p > 0.05$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

scored lower in the cytokine-associated hallmark pathway, indicating an underlying immunosuppressive microenvironment (Figure 2H).

Identified subtypes exhibited specific tumor microenvironments

The next step is to extensively focus on the microenvironmental components across tumor subtypes. The C_{EB1} subtype exhibited a less favorable immune score compared to C_{oriLyt} and C_{neg} subtypes, with higher tumor purity, validated in both internal and external cohorts (Figures 3A–B, S1A). Tumor Immune Estimation Resource (TIMER) method evaluation indicated significantly lower proportions of immune cells, including B cells, CD4+ and CD8+ T cells, and macrophages, in C_{EB1} subtype patients (Figures 3C, S1B). Using ssGSEA, the validation set showed minimal immune cell infiltration in the C_{EB1} subtype, particularly for B cells and CD4+/CD8+ T cells, compared to the C_{oriLyt} subtype with the highest immune cell expression (Figure 3D). Further assessment on the expression of chemokines and interleukin family members indicated that the C_{EB1} subtype exhibited markedly lower expression levels compared to the other two subtypes (Figure 3E). Additionally, downregulation of most cytotoxicity genes, immune checkpoints, and major histocompatibility complex (MHC) genes was observed in C_{EB1} subtype, contrasting with upregulation in C_{oriLyt} and C_{neg} subtypes (Figures 3F–H). These findings suggest potential limited response to immune checkpoint inhibitor therapy in C_{EB1} subtype patients.

Establishing and validating the prognostic signature for NPC

We created a risk signature focusing on the pivotal genes of the identified subtypes through WGCNA applied to gene expression data. The genes in the green, red, and turquoise modules corresponded to the C_{oriLyt}, C_{neg}, and C_{EB1} subtypes, respectively (Figure 4A). Key genes with prognostic significance were chosen from these modules (Figure 4B). From the chosen genes, 14 genes were validated and selected for the prognostic model via LASSO regression (Figure S2A), leading to the development of a risk score model. Subsequently, a risk score model was formulated with the equation: Risk score = $0.1515 \times$ BMPER + $0.1719 \times$ SPSB4 + $0.3283 \times$ SLAMF9 − $0.5385 \times$ CLEC4E + $0.0014 \times$ DKK1 + $0.3454 \times$ IGSF1 + $1.0983 \times$ RIMS2 + $0.0056 \times$ SPP1 + $0.0703 \times$ PTX3 + $0.3797 \times$ CD276 + $0.2150 \times$ BCHE + $0.0894 \times$ BMP2 (Table S4). Next, we examined the link between survival status and risk score. Results showed fewer surviving patients in the high-risk group compared to the low-risk group (Figure S2B). Kaplan-Meier analysis confirmed that the high-risk score was related to worse disease-free survival (DFS) in Fujian cancer hospital cohort (Figure 4C). Furthermore, we validated the prognostic significance of the risk model within the subgroup analyses. A higher risk score was linked to poorer DFS in the C_{oriLyt}, C_{neg}, and C_{EB1} subtypes (Figure S2C). This result was validated in the validation cohort (Figure S3A).

The NPC risk signature demonstrates robust prognostic assessment capabilities

The risk model of NPC exhibited substantial predictive capability in prognostic evaluation. To assess our risk model's performance, we evaluated its predictive accuracy for one-, three-, and five-year survival using receiver operating characteristic (ROC) curve analysis, yielding areas under the curve of 0.79, 0.86, and 0.88 (all > 0.7), respectively (Figure 4D). The external validation set, GSE102349, also exhibited strong predictive power, with respective area under curve (AUC) value for 2-year, 3-year, and 4-year survival outcomes of 0.73, 0.70, and 0.77, as illustrated in Figure S3B. Compared to classical clinical features like age, gender, TNM stage, and clinical stage, this risk signature demonstrated superior predictive efficiency and stable AUC values between 0.8 and 0.9 (Figure 4E). The risk score within the validation set escalated in correlation with the advancing clinical stages, indicating that an increased risk score is a predictive marker of disease advancement (Figure S3C). Univariate Cox regression analyses confirmed the predicted power of the risk score for DFS, revealing a significant association between high-risk scores and poorer DFS outcomes (Figure 4F). Multivariate analyses further corroborated that a high-risk score remained an independent predictor of worse DFS, even after accounting for other clinical variables (Figure 4G).

Predictive power of the efficacy of conventional treatment and immunotherapy

Limited responsiveness to diverse therapeutic strategies frequently characterizes poor prognostic outcomes. Building on our prior findings linking this risk score with adverse prognosis, we next examined its utility as a predictive biomarker for therapeutic response to chemotherapy, radiotherapy and immunotherapy. Chemotherapeutic agents utilized in the management of NPC, such as docetaxel and paclitaxel, encounter resistance issues in high-risk patients, given that their efficacy is inversely associated with the risk score (Figure 5A). A high-risk score is associated with elevated radioresistance scores (Figure 5B), indicating potential insensitivity to radiotherapy. We also observed high scores of radiotherapy resistance in the C_{EB1} subtype (Figure 5C). With standard treatment showing insensitivity in these cases, attention turned towards assessing immunotherapy response in patients with high-risk scores. Based on the tumor immune dysfunction and exclusion (TIDE) results, low IFNγ expression in the C_{EB1} subtype was indicative of poor immunotherapy response (Figure 5D). A smaller proportion of patients in the high-risk group exhibited a response to immunotherapy compared to those in the low-risk group (Figure 5E). Moreover, we discovered a negative correlation between risk score and im-

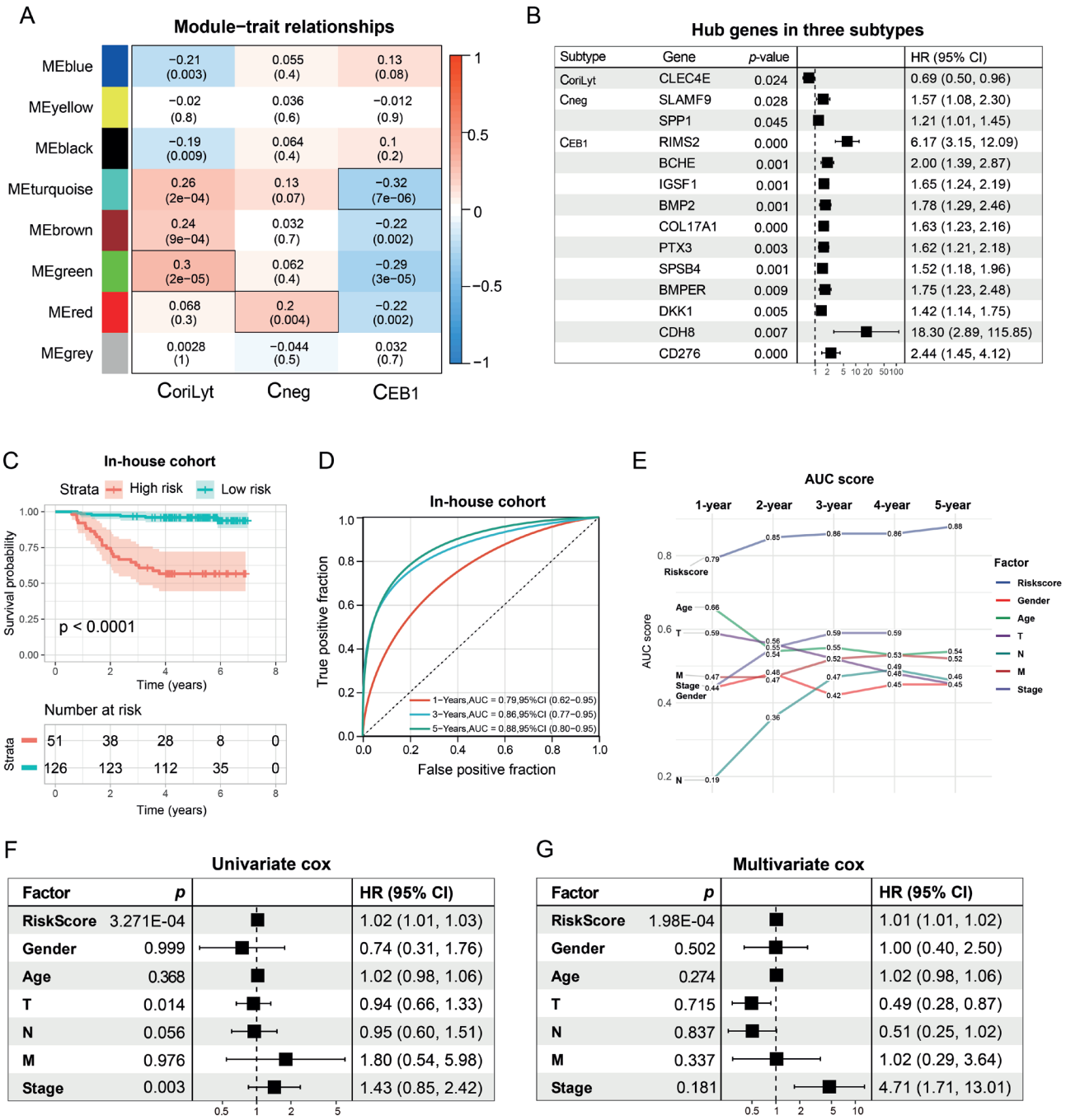


Figure 4. Establishment and verification of the NPC prognostic signature with a strong power for prognosis assessment. (A) Heatmap of the correlation between module eigengenes and subtypes of NPC. Each table cell contains the correlation coefficient and p-value. The color shade represents the correlation coefficient and p-value is described in parenthesis; (B) Univariate Cox analysis of key genes identified by WGCNA in three subtypes; (C) Kaplan–Meier curves for patients with high- or low-risk scores in the in-house training cohort. DFS is selected as a statistical indicator; (D) ROC curve showing the predictive value of NPC risk signature for 1-, 3-, and 5-year survival rates; (E) Comparison of predictive value between NPC risk signature and clinicopathologic features; (F–G) Univariate Cox (F) and multivariate Cox analyses (G) evaluating the independent prognostic value of the NPC risk signature in terms of DFS.

mune checkpoint expression in our internal cohort, which was validated in the GSE102349 dataset (Figures 5F–G).

M2 macrophages drive high-risk immune profiles and predict poor prognosis in viral-associated tumor environments

The NPC immune microenvironment is both complex and dynamic, with EBV infection driving active immune cell engagement. To elucidate the underlying pathological mechanisms, we

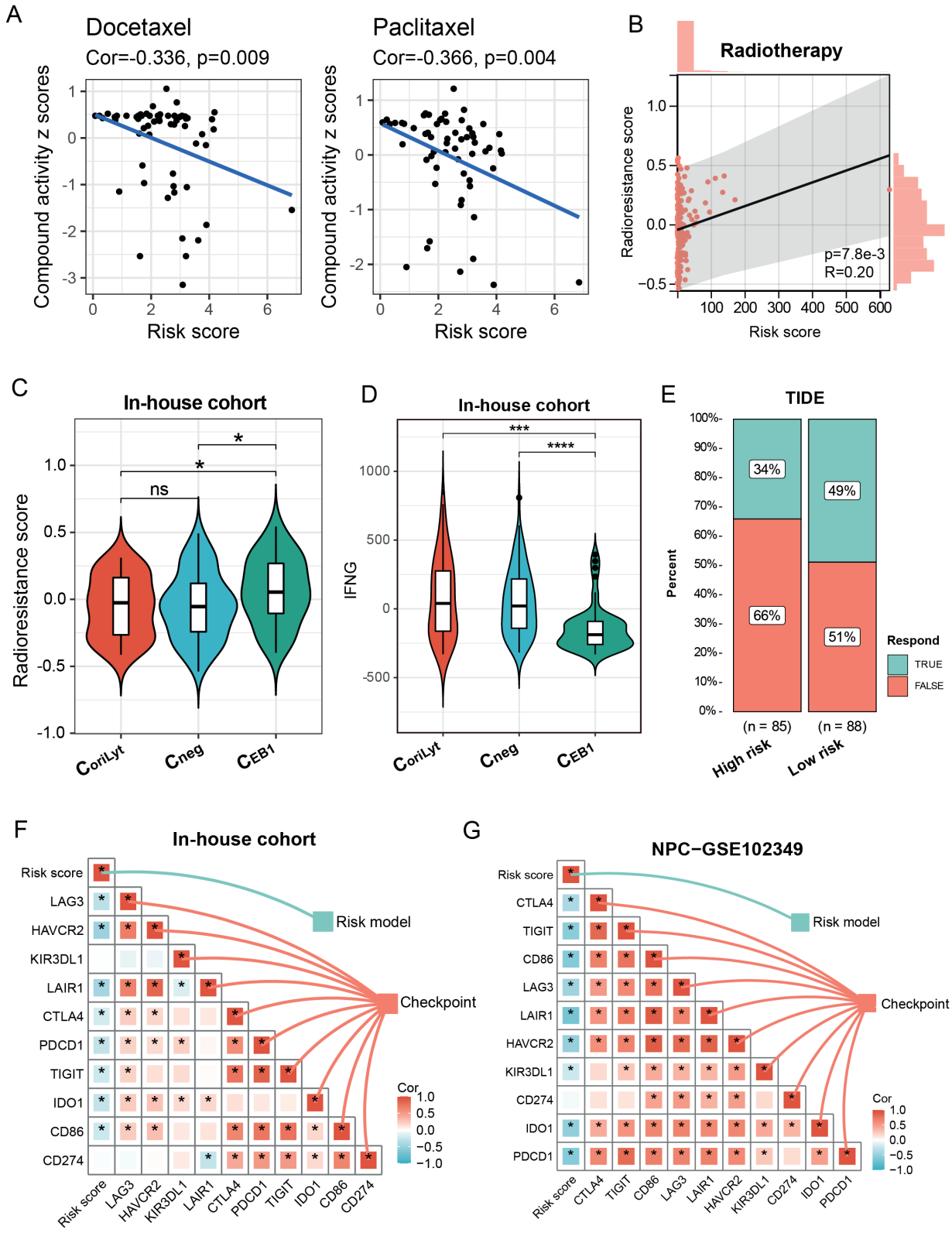


Figure 5. Risk model predicts the response of chemotherapy, radiotherapy and immunotherapy. (A) The relationship between risk score and chemotherapy drug sensitivity was evaluated; (B) The correlation between risk score and radiation therapy resistance scores; (C) The difference of radiation therapy resistance scores among C_{oriLyt}, C_{neg}, and C_{EB1} subtypes; (D) The difference of IFNG expression level among C_{oriLyt}, C_{neg}, and C_{EB1} subtypes; (E) Percentage of patients in high and low risk groups who may respond or may not respond to immunotherapy; (F–G) The correlation between risk score and immune checkpoint expression in the in-house cohort (F) and GSE102349 cohort (G).

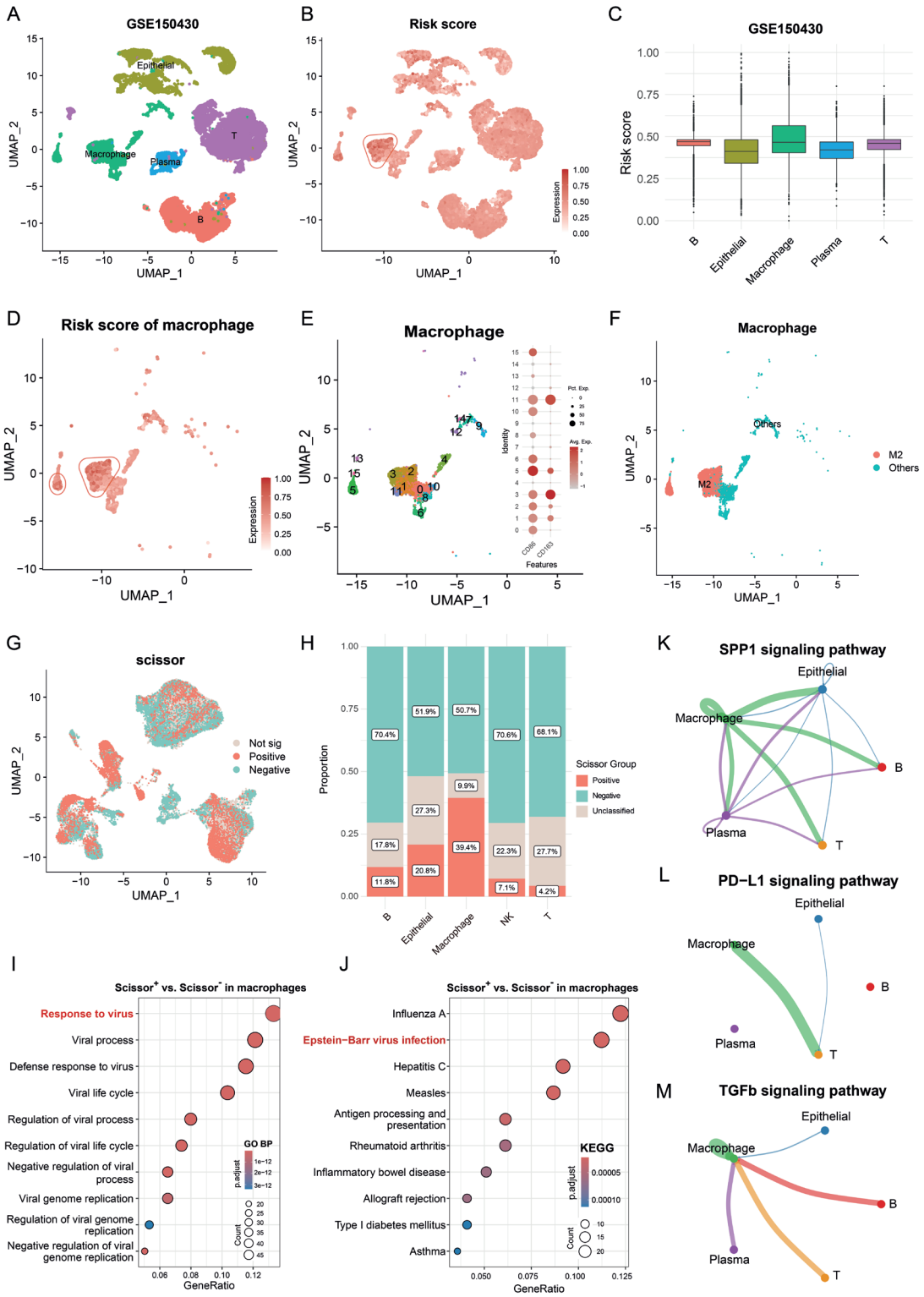


Figure 6. High-risk profile and prognostic impact of M2 macrophages in viral immune response. (A) UMAP projection of single-cell dataset GSE150430, with each color denoting a distinct cell type; (B) Mapping of risk score values at the single-cell level, with high-risk regions circled in red; (C) Boxplot illustrating quantitative risk score values across different cell populations; (D) UMAP mapping of risk scores within macrophages; (E) Subtype analysis of macrophages with corresponding marker expression levels; (F) Annotation of macrophage subpopulations; (G) Scissor algorithm results identifying cell populations positively or negatively correlated with disease-free survival (DFS); gray indicates no correlation, red signifies positive correlation with poor DFS, and blue indicates negative correlation with poor DFS; (H) Proportion analysis of cell populations identified by Scissor with significant DFS associations; (I-J) GO biological process (I) and KEGG pathway (J) enrichment analyses of differentially expressed genes between Scissor-positive and Scissor-negative macrophages; (K-M) Intercellular communication signals between macrophages and other cell subpopulations.

examined the viral immunity landscape at a single-cell resolution. Our analysis of 16 NPC samples identified five distinct cell subpopulations (Figures S4A–B, Figure 6A), highlighting an intricate interplay between tumor and immune cells. Risk score mapping revealed macrophages as the highest-risk cell population, positioning them as potential key mediators in protumorigenic activity (Figures 6B–C). Given the diversity within macrophage subtypes, further annotation identified the M2 subtype as contributing the highest risk score proportion (Figures 6D–F). To investigate the link between macrophages and poor prognosis, we employed the Scissor algorithm, integrating bulk and single-cell data, which identified cell populations correlating positively or negatively with patient prognosis (Figure 6G). Analysis confirmed macrophages as the predominant cell type in populations associated with poor outcomes (Figure 6H), supporting our initial findings. Enrichment analysis of DEGs in scissor+ and scissor- macrophages linked to prognosis revealed significant enrichment in viral response pathways (GO biological process, Figure 6I) and EBV infection pathways (KEGG, Figure 6J), underscoring their distinct role in EBV-driven immunity. Understanding macrophage signaling pathways is essential; CellChat analysis demonstrated active signaling networks among macrophages (Figures 6K–M, Figures S4C–F). The active SPP1 signaling pathway as a malignant feature underscores their malignancy-promoting activities (Figure 6K), while the active PD-L1 pathway highlights immune-suppressive signaling (Figure 6L). Additionally, TGF-β signaling—known to regulate immune overactivation—was primarily emitted by macrophages, further affirming their immunosuppressive role (Figure 6M).

Discussion

EBV, a known high-risk factor, is intricately linked to the onset of NPC. Precision oncology leverages sophisticated molecular profiling methods to pinpoint excellent biomarkers within tumors. Integrating an understanding of disease progression with risk factors, along with precision oncology strategies, offers novel perspectives in cancer diagnostics and personalized therapies. In our research (Figure S5), we conducted a novel molecular classification of NPC through transcriptome profiling of re-annotated EBV sequences, which were categorized into three distinct clinical subtypes: C_{oriLyt}, C_{neg}, C_{EB1}. Our findings indicate that cluster C_{EB1} is correlated with more advanced clinical stages and a poorer long-term prognosis, characterizing an immunosuppressive tumor microenvironment. In contrast, C_{oriLyt} and C_{neg} subtypes predominantly feature earlier clinical stages and a more favorable prognosis. We employed WGCNA to pinpoint pivotal genes for each cluster. Additionally, we created and verified a prognostic model using these key genes, demonstrating significant potential for enhanced prognostic evaluation. Our findings may inform risk-adapted therapeutic strategies for NPC patients. The identification of molecular subtypes, especially the immunosuppressive C_{EB1} subtype, suggests potential benefits from alternative immunotherapeutic combinations targeting the TGF-β or TNF-II pathways. Moreover, our subtype-derived risk model demonstrated strong prognostic value, which could aid in tailoring therapeutic intensity. High-risk patients may require more aggressive adjuvant therapies and closer surveillance schedules, while low-risk patients might be spared from overtreatment. Integration of this risk model into clinical workflows may enhance personalized care in NPC.

With regard to genomic studies, a variety of genetic alterations have been identified in NPC, including amplification of the CCND1 gene, mutation of the TP53 gene, and activation of cancerous signaling pathways [3,22–25]. Recent classification schemes for NPC based on miRNA expression, DNA methylation, or host transcriptomic profiles have provided important molecular insights but often lacked clear biological interpretation or clinical utility. For example, miRNA-based subtypes defined in earlier studies lacked consistent immune or metabolic correlates [13]. In addition, genome sequencing has linked specific EBV subtypes to an elevated risk of nasopharyngeal cancer, but the impact on patient stratification remains understated. In contrast, we reannotated the EBV viral RNA expression profiles of 193 patients with NPC, revealing three EBV-associated subtypes (C_{oriLyt}, C_{neg}, and C_{EB1}) with distinct immune, metabolic, and prognostic features. Notably, our subtypes capture EBV-driven biological variation and provide prognostically significant groups, thereby bridging viral etiology with clinical relevance—an aspect not addressed in prior systems. However, it's essential to validate our findings in an independent internal cohort, which should be considered when interpreting the study results.

Additionally, we discovered a set of key genes within these subtypes that strongly correlated with NPC prognosis. BMPER, essential for full activation of bone morphogenetic proteins signaling, is highly expressed in malignant tumors and critical for tumor growth [26]. Similarly, CLEC4E is significantly upregulated in gastric cancer, where its high expression is linked to poor prognosis and enhances cancer cell migration and invasion [27]. DKK1 contributes to tumor immune evasion and resistance to anti-PD-1 therapy in gastric cancer by recruiting immunosuppressive macrophages [28]. In non-small cell lung cancer, IGSF1 is more expressed in cells with low PD-L1 expression, while BMP2 overexpression is tightly associated with advanced tumor stages and increased metastatic load [29,30]. In colorectal cancer, RIMS2 is hypermethylated and underexpressed, affecting patient prognosis [31]. Conversely, tumor suppressor genes like SPSB4 are associated with the suppression of specific miRNAs in the tumor microenvironment of colon cancer and are elevated in TME cells [32]. SLAMF9, upregulated in melanoma, inhibits cell migration and has immunomodulatory effects on macrophages [33]. SPP1 plays a crucial role in determining tumor-associated macrophage polarity and serves as a prognostic indicator [34]. Defects

in PTX3 enhance autophagy in gliomas, which is key to controlling ferritin breakdown [35]. CD276 helps cancer stem cells evade the immune system in head and neck squamous cell carcinoma, suggesting that targeting it may reduce their numbers [36]. Lastly, BCHE is a key prognostic factor in endometrial cancer, with a negative association with CD4+ regulatory T cells [37]. However, it is necessary to conduct additional validation experiments either in vitro or in vivo to fully elucidate the functions of these key genes.

Immune checkpoint inhibitors (ICIs), a cornerstone of immunotherapy, have shown remarkable efficacy in the treatment of NPC [38]. Despite this, the therapeutic response has been variable, with some patients ultimately developing resistance to ICIs. Stratifying patients into high and low susceptibility groups could enhance the precision and effectiveness of immunotherapy. While various predictive biomarkers such as tumor mutational burden (TMB), microsatellite instability (MSI), lymphocyte infiltration, and immune scores, have been proposed, they individually offer limited predictive power [39]. Beyond subtype identification and prognostic stratification, our findings may also have implications for current therapeutic approaches in NPC. The C_{EB1} subtype, for instance, displayed features of immune suppression and metabolic activation, suggesting potential resistance to standard immunotherapies such as PD-1/PD-L1 blockade. These patients might instead benefit from combination strategies that co-target immunosuppressive pathways (e.g., TGF-β, TNF-II) or metabolic vulnerabilities. In contrast, C_{oriLyt} and C_{neg} subtypes exhibited higher immune infiltration and less pronounced metabolic signatures, which may render them more responsive to immune checkpoint inhibitors or conventional chemoradiotherapy. Therefore, our molecular subtypes could help guide treatment sensitivity predictions and enable more refined, risk-adapted therapeutic decision-making in clinical settings.

Limitations

This study has several limitations. First, although our cohort was relatively large and clinically annotated, it was derived from a single institution, which may limit the generalizability of the findings. Second, the external validation cohort lacked detailed clinical outcomes and EBV load data, which restricted further validation of subtype-specific prognostic features. Third, functional validation of the identified molecular subtypes and their therapeutic implications was not conducted in vitro or in vivo. Lastly, potential HPV co-infection, particularly in EBV-negative or keratinizing subtypes, was not assessed and warrants further investigation. Future studies exploring HPV-associated molecular features could further expand our understanding of NPC heterogeneity.

Conclusion

This study introduces an innovative transcriptomic-based classification system for NPC, utilizing EBV gene expression patterns. This classification holds significant promise in prognosticating the survival outcomes of patients with NPC.

Conflict of interest

The authors declare no competing interests.

Funding

This work was supported by the grants of Fujian Provincial Clinical Research Center for Cancer Radiotherapy and Immunotherapy (2020Y2012); Supported by the National Clinical Key Specialty Construction Program (2021); Fujian Clinical Research Center for Radiation and Therapy of Digestive, Respiratory and Genitourinary Malignancies (2021Y2014). National Natural Science Foundation of China (82473376, 12374405); Major Scientific Research Program for Young and Middle-aged Health Professionals of Fujian Province, China (2021ZQNZD010); Joint Funds for the Innovation of Science and Technology, Fujian province (2021Y9196), Natural Science Foundation of Fujian Province (2023J011267,2024J011108, 2024J011086), and High-level Talent Training Program of Fujian Cancer Hospital (2022YNG07), Subsidy for Young and Middle-aged Experts with Outstanding Contributions in Fujian Province's Health System in 2021-2022 (F23R-TG01-01).

Authors contributions

Conceptualization: QD; Methodology: QD, YP; Investigation: WL, HY; Writing - Original Draft: QD, WL; Writing, review & editing: XC, HL; Funding acquisition: SQ; Resources: SQ; Supervision: SQ, YW.

Acknowledgement

Not applicable.

Ethics approval

This study was authorized by the ethics committee of Fujian Cancer Hospital (Fuzhou, China; numbers K2022-074-01). Each patient was asked to grant their written and informed consent before participating in any study-specific research.

Data availability

All original contributions discussed in this study are included in the article and its Supplementary Material. The data presented in this study are available in the GEO repository under accession number GSE1150430 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150430). Additional information will be provided by the authors upon request, without any undue restrictions.

References

1. Tang LL, Chen YP, Chen CB, et al. The Chinese Society of Clinical Oncology (CSCO) clinical guidelines for the diagnosis and treatment of nasopharyngeal carcinoma. Cancer Commun (Lond). 2021 Nov;41(11):1195–227.
2. Lee AWM, Ng WT, Chan JYW, et al. Management of locally recurrent nasopharyngeal carcinoma. Cancer Treat Rev. 2019 Sep;79:101890.
3. Chen YP, Chan ATC, Le QT, Blanchard P, Sun Y, Ma J. Nasopharyngeal carcinoma. Lancet. 2019 Jul 6;394(10192):64–80.
4. Jen C, Tsai Y, Wu JS, et al. Prognostic classification for patients with nasopharyngeal carcinoma based on American Joint Committee on cancer staging system T and N categories. 2020;
5. Bossi P, Chan AT, Licitra L, Trama A, Orlandi E, Hui EP, et al. Nasopharyngeal carcinoma: ESMO-EURACAN clinical practice guidelines for diagnosis, treatment and follow-up†. Ann Oncol. 2021 Apr;32(4):452–65.
6. Guo R, Mao YP, Tang LL, Chen L, Sun Y, Ma J. The evolution of nasopharyngeal carcinoma staging. Br J Radiol. 2019 Oct;92(1102):20190244.
7. Trevisiol C, Gion M, Vaona A, et al. The appropriate use of circulating EBV-DNA in nasopharyngeal carcinoma: comprehensive clinical practice guidelines evaluation. Oral Oncol. 2021 Mar;114:105128.
8. Lee AWM, Lee VHF, Ng WT, et al. A systematic review and recommendations on the use of plasma EBV DNA for nasopharyngeal carcinoma. Eur J Cancer. 2021 Aug;153:109–22.
9. Ding RB, Chen P, Rajendran BK, et al. Molecular landscape and subtype-specific therapeutic response of nasopharyngeal carcinoma revealed by integrative pharmacogenomics. Nat Commun. 2021 May 24;12(1):3046.
10. Argirion I, Zarins KR, Suwanrungruang K, et al. Subtype specific nasopharyngeal carcinoma incidence and survival trends: differences between endemic and non-endemic populations. Asian Pac J Cancer Prev. 2020 Nov 1;21(11):3291–9.
11. Mardis ER. The impact of next-generation sequencing on cancer genomics: from discovery to clinic. Cold Spring Harb Perspect Med. 2019 Sep 3;9(9):a036269.
12. MacConaill LE, Van Hummelen P, Meyerson M, Hahn WC. Clinical implementation of comprehensive strategies to characterize cancer genomes: opportunities and challenges. Cancer Discov. 2011 Sep;1(4):297–311.
13. Zhao L, Fong AHW, Liu N, Cho WCS. Molecular subtyping of nasopharyngeal carcinoma (NPC) and a microRNA-based prognostic model for distant metastasis. J Biomed Sci. 2018 Feb 19;25(1):16.
14. Ka-Yue Chow L, Lai-Shun Chung D, Tao L, et al. Epigenomic landscape study reveals molecular subtypes and EBV-associated regulatory epigenome reprogramming in nasopharyngeal carcinoma. EBioMedicine. 2022 Dec;86:104357.
15. Lin W, Chen X, Huang Z, Ding Q, Yang H, Li Y, et al. Identification of novel molecular subtypes to improve the classification framework of nasopharyngeal carcinoma. Br J Cancer. 2024 Jan 27;
16. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015 Jan 15;31(2):166–9.
17. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. Imeta. 2023;2(2):e107.
18. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016 Sep;11(9):1650–67.
19. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008 Dec 29;9:559.
20. Oldham MC, Konopka G, Iwamoto K, et al. Functional organization of the transcriptome in human brain. Nat Neurosci. 2008 Nov;11(11):1271–82.
21. Voineagu I, Wang X, Johnston P, et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 2011 May 25;474(7351):380–4.
22. Li YY, Chung GTY, Lui VWY, et al. Exome and genome sequencing of nasopharynx cancer identifies NF-κB pathway activating mutations. Nat Commun. 2017 Jan 18;8:14121.
23. Zhang L, MacIsaac KD, Zhou T, et al. Genomic analysis of nasopharyngeal carcinoma reveals TME-based subtypes. Mol Cancer Res. 2017 Dec;15(12):1722–32.
24. Lin DC, Meng X, Hazawa M, et al. The genomic landscape of nasopharyngeal carcinoma. Nat Genet. 2014 Aug;46(8):866–71.
25. Zheng H, Dai W, Cheung AKL, et al. Whole-exome sequencing identifies multiple loss-of-function mutations of NF-κB pathway regulators in nasopharyngeal carcinoma. Proc Natl Acad Sci U S A. 2016 Oct 4;113(40):11283–8.
26. Heinke J, Kerber M, Rahner S, et al. Bone morphogenetic protein modulator BMPER is highly expressed in malignant tumors and controls invasive cell behavior. Oncogene. 2012 Jun 14;31(24):2919–30.
27. Jiang Q, Xiao D, Wang A, et al. CLEC4E upregulation in gastric cancer: a potential therapeutic target correlating with tumor-associated macrophages. Heliyon. 2024 Mar 15;10(5):e27172.
28. Shi T, Zhang Y, Wang Y, et al. DKK1 promotes tumor immune evasion and impedes Anti-PD-1 treatment by inducing immunosuppressive macrophages in gastric cancer. Cancer Immunol Res. 2022 Dec 2;10(12):1506–24.
29. Wu CK, Wei MT, Wu HC, Wu CL, Wu CJ, Liaw H, et al. BMP2 promotes lung adenocarcinoma metastasis through BMP receptor 2-mediated SMAD1/5 activation. Sci Rep. 2022 Sep 29;12(1):16310.
30. Koh DI, Lee M, Park YS, et al. The immune suppressor IGSF1 as a potential target for cancer immunotherapy. Cancer Immunol Res. 2024 Apr 2;12(4):491–507.
31. Li X, Cai D, Huang Y, et al. Aberrant methylation in neurofunctional gene serves as a hallmark of tumorigenesis and progression in colorectal cancer. BMC Cancer. 2023 Apr 6;23(1):315.
32. Gerovska D, Garcia-Gallastegi P, Crende O, Márquez J, Larrinaga G, Unzurrunzaga M, et al. GeromiRs are downregulated in the tumor microenvironment during colon cancer colonization of the liver in a murine metastasis model. Int J Mol Sci. 2021 May 1;22(9):4819.
33. Dollt C, Michel J, Kloss L, et al. The novel immunoglobulin super family receptor SLAMF9 identified in TAM of murine and human melanoma influences pro-inflammatory cytokine secretion and migration. Cell Death Dis. 2018 Sep 19;9(10):939.
34. Su X, Liang C, Chen R, Duan S. Deciphering tumor microenvironment: CXCL9 and SPP1 as crucial determinants of tumor-associated macrophage polarity and prognostic indicators. Mol Cancer. 2024 Jan 13;23(1):13.
35. Lathoria K, Gowda P, Umdor SB, Patrick S, Suri V, Sen E. PRMT1 driven PTX3 regulates ferritinophagy in glioma. Autophagy. 2023 Jul;19(7):1997–2014.
36. Wang C, Li Y, Jia L, et al. CD276 expression enables squamous cell carcinoma stem cells to evade immune surveillance. Cell Stem Cell. 2021 Sep 2;28(9):1597-1613.e7.
37. Liu J, Tian T, Liu X, Cui Z. BCHE as a prognostic biomarker in endometrial cancer and its correlation with immunity. J Immunol Res. 2022;2022:6051092.
38. Wang FH, Wei XL, Feng J, et al. Efficacy, safety, and correlative biomarkers of toripalimab in previously treated recurrent or metastatic nasopharyngeal carcinoma: a phase ii clinical trial (POLARIS-02). J Clin Oncol. 2021 Mar 1;39(7):704–12.
39. Gavrielatou N, Doumas S, Economopoulou P, Foukas PG, Psyrri A. Biomarkers for immunotherapy response in head and neck cancer. Cancer Treat Rev. 2020 Mar;84:101977.

Sufang Qiu
Department of Radiation Oncology
Clinical Oncology School of Fujian
Medical University
Fujian Cancer Hospital
Fuzhou 350014
China

E-mail: sufangqiu@fjmu.edu.cn

Youliang Weng
Department of Radiation Oncology
Clinical Oncology School of Fujian
Medical University
Fujian Cancer Hospital
Fuzhou 350014
China

E-mail: wyl7788@sina.com

Qin Ding[1,2#], Yuhui Pan[1,2#], Wanzun Lin[1,2], Hanxuan Yang[1,2], Xin Chen[1,2], Haolan Li[1,2], Youliang Weng[1,2*], Sufang Qiu[1,2*]

[1] Clinical Oncology School of Fujian Medical University, Fujian Cancer Hospital (Fujian Branch of Fudan University Shanghai Cancer Center), Fuzhou, China

[2] Fujian Provincial Key Laboratory of Translational Cancer Medicine, Fuzhou, China

[#] These authors contributed equally to this work.

Rhinology 64: 1, 0 - 0, 2026
https://doi.org/10.4193/Rhin25.274

Received for publication:
May 23, 2025
Accepted: August 6, 2025

Associate Editor:
Basile Landis

This manuscript contains online supplementary material

SUPPLEMENTARY MATERIAL

Contents

Supplementary experimental procedures

Detecting differentially expressed genes (DEGs) between NPC and normal tissues

The "limma" package [1] was utilized to identify differentially expressed genes (DEGs) in tumor versus normal tissues, setting a threshold of |log2 fold change| (|logFC|) > 1.5 and an adjusted p-value (P.adj) < 0.05. The "ggplot2" package was employed to visualize the volcano plot of DEGs [2].

Gene set variation analysis

Gene Set Variation Analysis (GSVA) was conducted using the "GSVA" package in R software (version 4.3.2) to compute pathway scores for NPC samples on the basis of transcriptome data [3,4]. GSVA evaluates gene expression levels within predefined gene sets, generating an enrichment score that reflects the overall activity of specific gene sets in each sample. Kyoto Encyclopedia of Genes and Genomes (KEGG) gene sets were employed for finer resolution of functional signature variations across samples [5].

Gene Ontology (GO) enrichment analysis

We performed gene set functional enrichment utilizing GO annotations from the R package org.Hs.eg.db (v3.1.0) to map genes to the background set (6). We used the R package clusterProfiler (v3.14.3) for enrichment analysis, with gene set sizes ranging from 5 to 5000 (7). Significance was determined by $p < 0.05$ and false discovery rate (FDR) < 0.25.

Consensus clustering

Molecular subtypes were identified via consensus clustering using "ConcensusClusterPlus" in R software. Optimal clustering values (k = 2 to 10) were determined through 1,000 iterations to ensure result reproducibility and robustness.

Principal component analysis (PCA)

PCA was conducted to examine the transcriptional profiles across various clusters and to assess the agreement among them by examining the distribution of the principal components. This analysis was carried out with the "princomp" function from the R package "limma". Subsequently, the findings were illustrated utilizing the "ggplot2" package for visual representation.

Survival analysis

The prognostic outcomes of patients with nasopharyngeal carcinoma in different subsections were analysed using the "survival" software package. The survival outcome used was DFS to assess the prognostic predictive value of the proposed subtypes.

Evaluation of immune cell infiltration level

We applied the TIMER algorithm to assess the infiltration levels of several immune cell types, including macrophages, B cells, CD4 T cells, CD8 T cells, neutrophils, and dendritic cells (DCs), in NPC samples. Furthermore, the ESTIMATE algorithm, implemented in R v4.3.2, was employed to estimate immune and tumor purity scores by analyzing the expression of a predefined gene set indicative of immune cell presence in tumor microenvironment. The immune infiltration status of samples can be calculated using the ssGSEA algorithm provided in the R package GSVA, which employs markers for 24 types of immune cells [8,9]. This approach allows for the quantification of immune cell fractions and the evaluation of their impact on tumor biology.

Constructing and validating the prognostic risk signature

Here the Least Absolute Shrinkage and Selection Operator (LASSO) Cox regression method was employed to determine the genes and their corresponding coefficient values within the risk model. LASSO is an analytical approach that enhances model predictability and interpretability by selecting variables and applying regularization. It is particularly adept at developing prognostic models from gene expression data, as evidenced by references [10–14]. GSE102349 was utilized as the validation set [15].

Chemotherapy and radiotherapy sensitivity evaluation

To assess risk models' relationship of commonly used drug sensitivity in NPC, we studied the NCI-60 cell line. We obtained drug sensitivity data, inhibitory concentration (IC50) values, from the CellMiner database [16]. We then analyzed 218 FDA-approved drugs and 574 drugs/compounds from trials. The impact of risk on drug sensitivity was evaluated using "impute" [17] and the "limma" [18] R package.

We investigated the effect of the risk model on radiotherapy sensitivity by evaluating radiotherapy tolerance in the internal cohort sample using the GSVA technique [9], with scores for each

sample calculated using the ssGSEA method in R [19].

Immunotherapy response prediction

To assess the predictive performance of the risk signature for immunotherapy response, we collected immunotherapy cohorts from the GEO database and the TIGER website (http://tiger.canceromics.org/#/). This included our in-house cohort. We visually compared response proportions in high- and low- risk groups. In addition, the correlation between immune checkpoint expression and risk score was computed to evaluate the treatment response with immune checkpoint inhibitors among patients with varying risk profiles.

References

1. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:Article3.
2. Galensa K. ggplot2: elegant graphics for data analysis (2nd ed.). 2017;58(8):457–8.
3. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013 Jan 16;14:7.
4. Becht E, Giraldo NA, Lacroix L, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016 Oct 20;17(1):218.
5. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27–30.
6. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25–9.
7. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012 May;16(5):284–7.
8. Bindea G, Mlecnik B, Tosolini M, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity. 2013 Oct 17;39(4):782–95.
9. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013 Jan 16;14:7.
10. Wei JH, Feng ZH, Cao Y, et al. Predictive value of single-nucleotide polymorphism signature for recurrence in localised renal cell carcinoma: a retrospective analysis and multicentre validation study. Lancet Oncol. 2019 Apr;20(4):591–600.
11. Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Statistical Society, Series B. 1996;58(1).
12. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med. 1997 Feb 28;16(4):385–95.
13. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med. 2007 Dec 30;26(30):5512–28.
14. Bøvelstad HM, Nygård S, Størvold HL, et al. Predicting survival from microarray data--a comparative study. Bioinformatics. 2007 Aug 15;23(16):2080–7.
15. Zhang L, MacIsaac KD, Zhou T, et al. Genomic analysis of nasopharyngeal carcinoma reveals TME-Based subtypes. Mol Cancer Res. 2017 Dec;15(12):1722–32.
16. Reinhold WC, Sunshine M, Liu H, et al. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. Cancer Res. 2012 Jul 15;72(14):3499–511.
17. Hastie T, Tibshirani R, Narasimhan B, Chu G. Impute: Imputation for microarray data. 2016 Jan 1;17:520–5.
18. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47.
19. Newman AM, Liu CL, Green MR, vgf gv et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015 May;12(5):453–7.

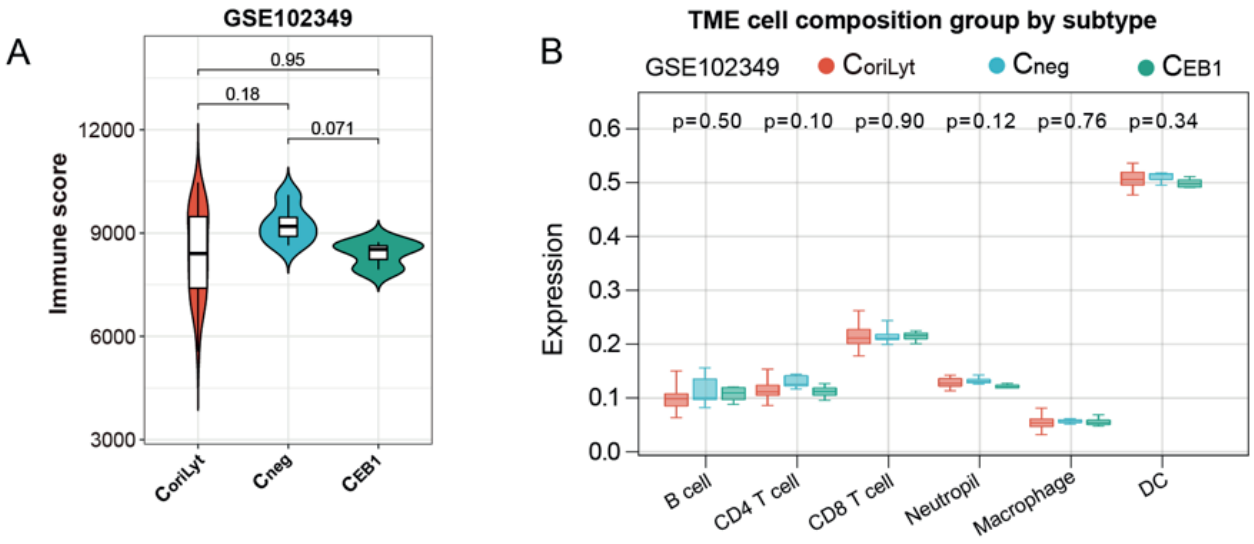


Figure S1. (A) Violin plots showing the immune score of GSE102349 cohort; (B) Box plot of 6 immune cell population score among three subtypes in GSE102349 validation cohort. Red boxes represent C_{oriLyt} subtype, blue boxes represent C_{neg} subtype, and green boxes represent C_{EB1} subtype.

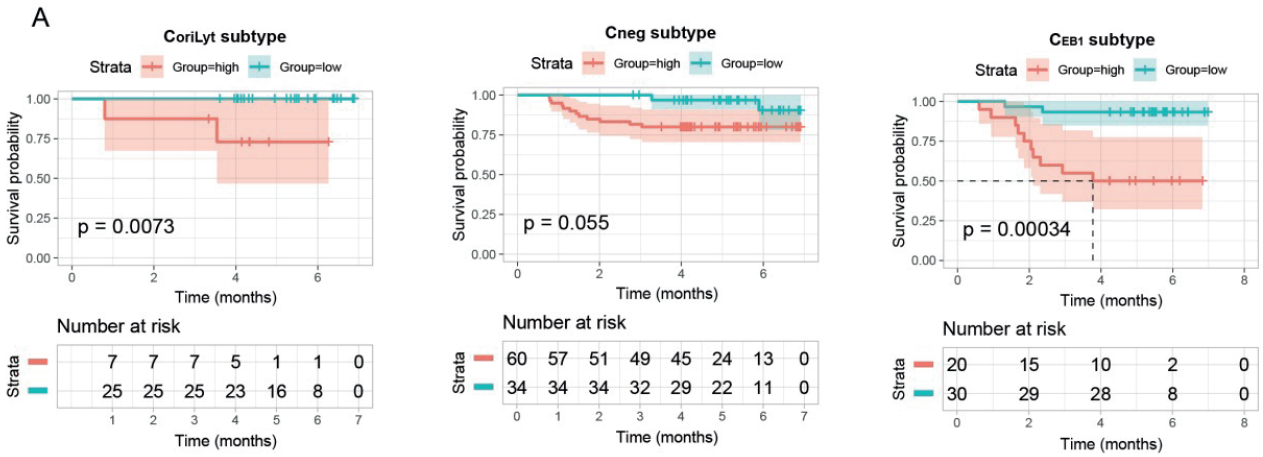


Figure S2. (A) The prognostic value of risk model in each subtype.

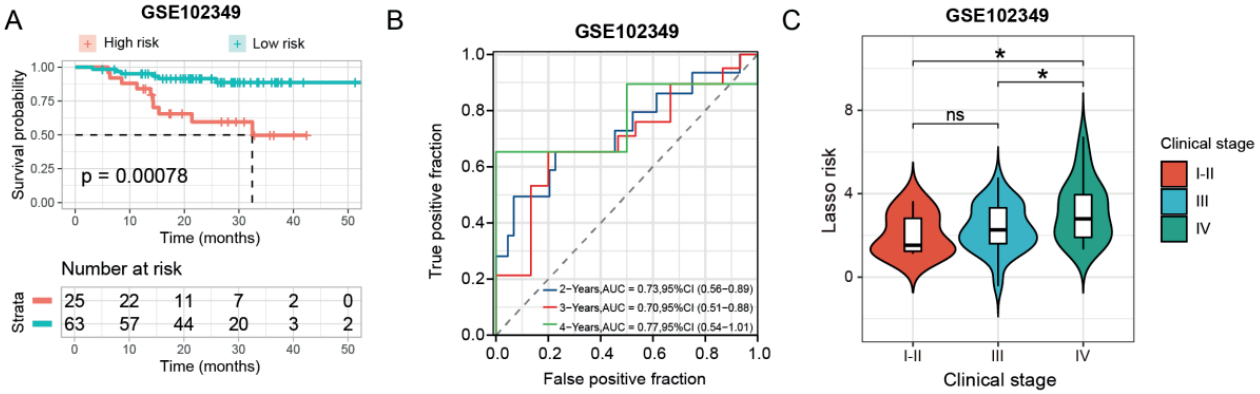


Figure S3. (A) Kaplan–Meier curves for patients with high- or low-risk scores in GSE102349 cohort. PFS is selected as a statistical indicator; (B) ROC curve showing the predictive value of NPC risk signature for 2-, 3-, and 4-year survival rates; (C) Violin plot showing risk scores between different clinical stages in the GSE102349 cohort. ns, $p > 0.05$; *, $p < 0.05$.

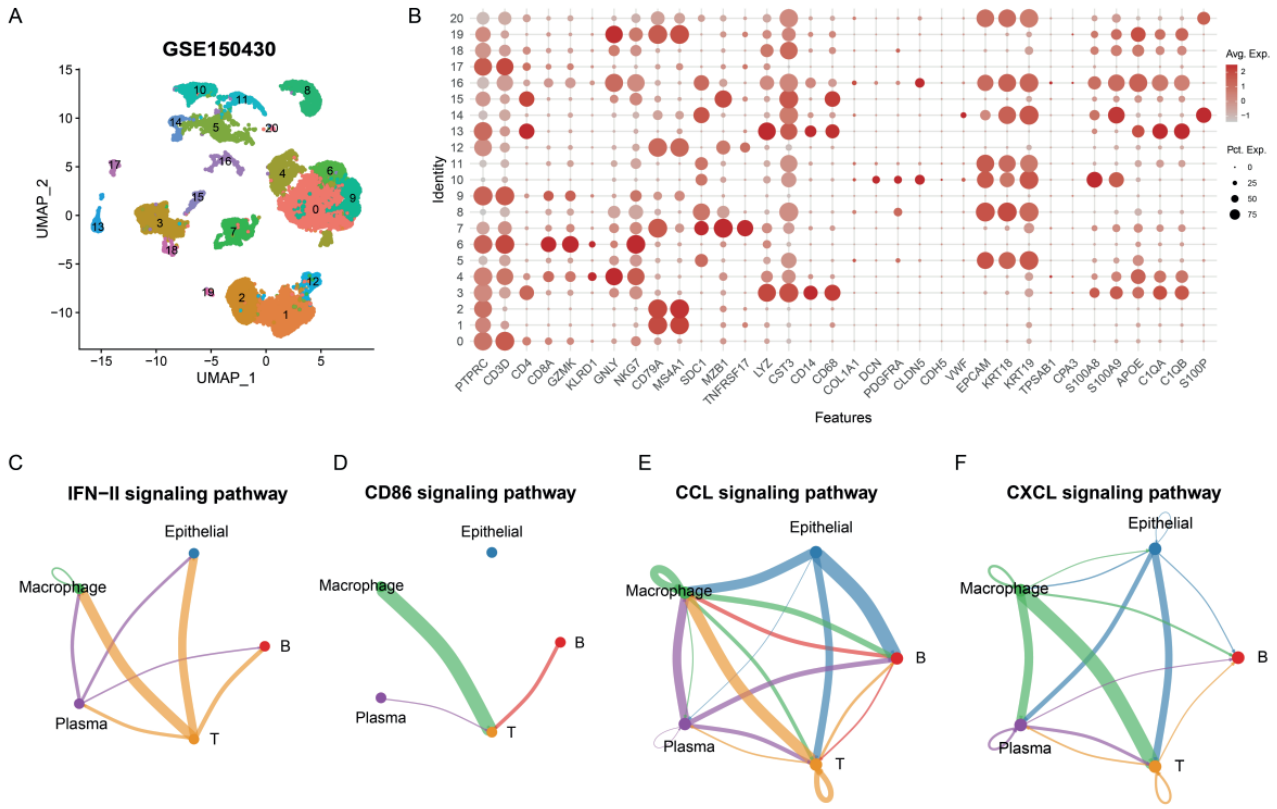


Figure S4. (A) UMAP projection of single-cell dataset GSE150430, with distinct colors representing various cell clusters; (B) Bubble plot displaying marker expression across cell clusters; (C-F) Intercellular communication signals between macrophages and other cell subpopulations

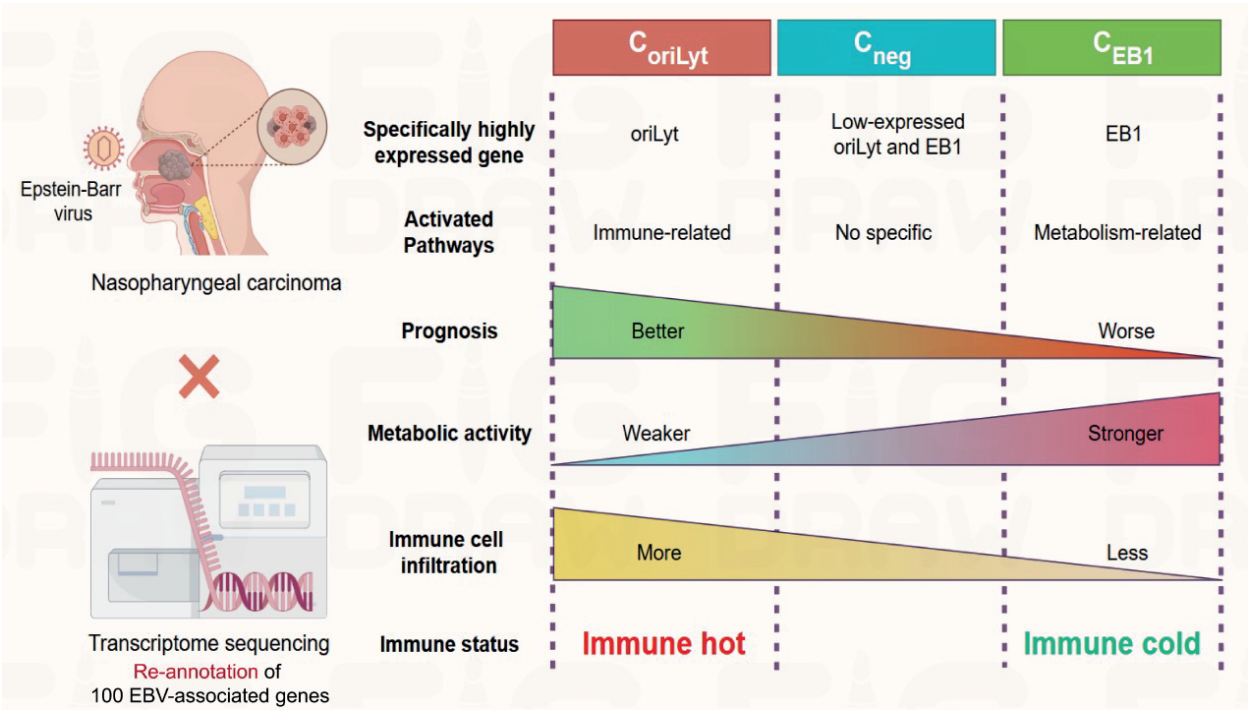


Figure S5. Summary of key findings. In this study, three distinct subtypes were identified through reannotation of transcriptomic RNA-seq data using EBV sequences. These subtypes exhibit unique characteristics in prognosis, metabolic activity, and immune microenvironment. Based on these findings, a robust risk model was constructed, demonstrating high reproducibility and predictive accuracy for prognosis and treatment response. Additionally, M2 macrophages closely associated with EBV infection were identified as active contributors in cases with poor prognosis.

Table S1. The sequencing coverage of each sample.

Sample Name	Total Clean Read	Total Mapping Genome Ratio	Uniquely Mapping Genome Ratio	Sample Name	Total Clean Read	Total Mapping Genome Ratio	Uniquely Mapping Genome Ratio
N1307A	23.87	94.34	72.13	T1013A	23.48	76.88	58.36
N1311A	23.66	94.06	74.75	T1016A	22.76	92.72	53.86
N1314A	23.68	94.76	71.22	T1020A	22.78	94.41	65.6
N1319A	23.88	94.89	73.1	T1023A	23.23	93.71	64
N1322A	20.18	94.55	73.05	T1024A	23.72	94.77	69.08
N1331A	19.9	94.91	67.69	T1026A	23.53	93.67	71.08
N1341A	23.43	90.68	66.7	T1056A	23.78	95.6	71.12
N1354A	22.31	82.07	61.79	T1058A	23.56	92.81	70.7
N200A	23.72	94.83	73.61	T1060A	20.64	86.85	66.28
N201A	23.74	94.6	69.51	T1061A	23.57	93.41	63.33
N202A	22.12	95.07	75.8	T1063A	23.78	93.27	55.24
N203A	23.73	94.61	68.95	T1067A	23.82	95.02	68.88
N204A	23.7	93.57	68.89	T1079A	23.84	93.68	57.38
N205A	23.77	94.53	69.78	T1082A	23.78	93.61	53.05
N206A	23.64	91.44	66.91	T1084A	23.08	92.25	69.41
N207A	21.68	95.17	75.63	T1085A	23.62	95.54	74.96
N208A	24.49	82.74	61.72	T1088A	15.21	79.61	57.36
N209A	23.43	90.64	67.4	T1089A	23.21	95.61	71.77
N210A	23.86	95.32	73.73	T1090A	25.37	86.96	65.45
N211A	21.73	89.71	59.98	T1091A	23.52	91.12	71.79
N213A	23.62	86.82	64.21	T1094A	23.15	89.64	64.96
N214A	23.43	87	61.33	T1096A	23.62	94.53	74.27
N215A	23.75	94.41	67.87	T1098A	23.77	94.9	68.15
N216A	23.47	88.13	66.82	T1104A	23.78	94.97	72.82
N217A	23.84	95.25	73.99	T1106A	23.72	95.44	71.77
N218A	22.72	87.14	62.35	T1108A	23.89	71.26	50.84
N219A	23.82	94.39	62.77	T1109A	23.66	94.58	71.67
N631A	23.77	95.46	75.32	T1112A	23.79	94.56	71.64
N633A	23.73	95.06	76.95	T1113A	23.62	94.4	69.19
N634A	23.45	95.46	73.36	T1114A	23.67	90.8	65.87
N635A	23.52	93.26	55.35	T1115A	23.78	94.62	72.08
N637A	23.69	94.16	66.13	T1119A	23.84	94.32	67.97
N638A	23.87	95.21	73.44	T1121A	23.78	94.43	73.52
N640A	23.86	93.03	70.54	T1123A	23.66	94.03	60.26
N641A	23.86	95.4	78.87	T1125A	22.98	91.09	68.62
N642A	23.86	94.13	70.16	T1127A	24.06	85.69	58.24
N649A	23.75	95.87	75.95	T1130A	23.73	92.1	64.37
N651A	23.87	95.59	76.87	T1136A	23.54	87.08	65.69
N654A	23.82	95.29	75.52	T1140A	23.67	95	69.97
T1000A	23.87	93.17	67.96	T1142A	23.65	94.39	71.94
T1005A	23.06	83	58.1	T1147A	21.03	89.78	67.16
T1007A	23.73	94.46	74.99	T1148A	23.59	94.27	68.64
T1008A	23.26	91.39	71.87	T1150A	23.71	95.06	71.9
T1010A	20.1	95.52	76.48	T1152A	23.29	88.6	66.38

Table S1 *continued*. The sequencing coverage of each sample.

Sample Name	Total Clean Read	Total Mapping Genome Ratio	Uniquely Mapping Genome Ratio
T1155A	23.63	93.99	63.27
T1156A	23.31	92.97	70.88
T1159A	23.48	94.83	74.07
T1161A	23.61	92.04	58.22
T1163A	19.63	76.87	57.46
T1164A	23.65	95.06	67.97
T1167A	23.37	93.76	65.69
T1168A	23.77	95.61	75.85
T1169A	23.24	83.53	62.47
T1172A	23.77	93.73	72.94
T1174A	23.68	90.66	61.45
T1179A	23.73	94.33	73.54
T1180A	21.09	92.96	60.33
T1181A	23.84	95.42	77.16
T1189A	23.06	94.49	70.22
T1190A	23.61	93.56	70.69
T1191A	23.72	94.32	70.57
T1193A	23.64	92.48	71.53
T1195A	23.54	93.39	63.21
T1199A	17.71	88.4	68.97
T1283A	23.43	94.47	66.37
T1284A	21.99	89.91	60.79
T1285A	23.5	92.75	74.06
T1286A	19.66	94.92	72.67
T1289A	23.46	93.93	64.1
T1291A	23.87	95.44	74.14
T1292A	23.39	94.23	68.67
T1295A	23.45	95.2	74.16
T1296A	23.39	94.92	74.82
T1297A	23.16	89.32	53.93
T1298A	22.82	93.65	72.57
T1299A	23.87	94.15	70.31
T1300A	23.45	93.73	73
T1301A	23.39	94.42	67.69
T1302A	23.49	94.74	69.76
T1303A	23.78	94.85	70.96
T1305A	23.44	94.14	68.37
T1306A	23.62	93.71	74.26
T1308A	23.53	92.38	68.05
T1309A	23.55	92.72	69.76
T1310A	23.54	92.52	69.29
T1313A	23.8	96.16	75.93
T1316A	21.32	93.73	74.79
T1317A	23.78	95.96	76.49

Sample Name	Total Clean Read	Total Mapping Genome Ratio	Uniquely Mapping Genome Ratio
T1318A	23.76	91.03	53.28
T1321A	19.64	93.43	70.03
T1323A	23.79	95.56	73.58
T1324A	20.15	89.48	70.19
T1325A	23.58	90.09	57.18
T1326A	23.81	94.33	74.09
T1327A	23.55	93.47	72.67
T1328A	23.58	92.53	67.47
T1329A	23.53	93.77	72.77
T1330A	21.97	92.08	61.8
T1333A	18.54	85.78	56.1
T1335A	23.87	94.83	73.97
T1337A	21.89	92.46	64.28
T1338A	23.78	95.52	72.36
T1339A	23.52	92.83	70.75
T1340A	23.55	92.74	70.44
T1342A	23.87	95.94	81.88
T1343A	23.8	95.87	74.74
T1346A	23.78	96.03	75.58
T1347A	23.78	95.06	73.8
T1349A	23.7	95.43	75.9
T1352A	23.81	94.78	73.02
T1355A	23.78	94.61	73.67
T636A	23.62	95.56	77.26
T639A	24.07	82.3	54.88
T648A	23.68	95.13	71.1
T653A	21.83	93.45	65.86
T658A	23.65	95.35	72.75
T667A	23.64	94.32	73.74
T670A	23.62	94.67	72.41
T685A	21.46	93.97	67.11
T751A	23.67	94.55	71.19
T754A	23.64	95.18	68.43
T757A	23.65	94.83	72.6
T763A	23.65	94.91	74.51
T764A	23.63	95.94	78.45
T774A	23.48	95.7	72.92
T775A	23.66	94.73	72.9
T785A	23.16	91.99	59.81
T806A	23.61	94.4	67.73
T810A	22.91	90.91	66.72
T820A	23.58	94.65	69.64
T821A	23.63	95.6	73.83
T823A	21.13	89.27	61.85

Table S1 *continued*. The sequencing coverage of each sample.

Sample Name	Total Clean Read	Total Mapping Genome Ratio	Uniquely Mapping Genome Ratio
T827A	23.59	95.77	71.18
T831A	23.53	94.06	75.38
T842A	23.58	94.06	69.3
T848A	23.53	94.97	73.14
T857A	23.62	93.85	64.28
T858A	23.22	90.17	69.19
T861A	23.38	94.34	64.64
T863A	23.65	94.63	71.25
T865A	23.58	95.8	68.81
T869A	23.29	93.78	67.98
T879A	22.44	92.94	70.07
T895A	23.63	94.26	70.56
T897A	23.63	94.67	72.29
T899A	23.62	94.85	73.32
T905A	23.6	94.88	74.18
T907A	24.05	94.7	69.71
T915A	23.64	95.23	72.49
T925A	20.16	88.52	70.16
T933A	23.74	95.31	66.69

Sample Name	Total Clean Read	Total Mapping Genome Ratio	Uniquely Mapping Genome Ratio
T935A	14.52	84.14	60.33
T943A	23.62	95.14	70.15
T944A	10.9	69.78	57.2
T948A	23.37	93.99	63.59
T951A	23.61	93.81	66.66
T953A	23.56	95.36	70.41
T956A	20.1	95.18	73.14
T957A	23.57	94.64	72.72
T959A	23.49	93.77	74.96
T960A	23.54	94.42	68.79
T961A	23.7	92.64	72.02
T967A	23.58	95.29	77.6
T972A	23.58	81.37	63.11
T974A	23.61	93.15	59.08
T977A	23.26	86.39	61.09
T984A	23.64	95	73.45
T986A	23.69	94.51	69
T997A	23.52	92.47	73.24

Table S2. The quality statistics of each sample.

Sample Name	Total Raw Reads	Total Clean Reads	Total Clean Bases	Clean Reads Q20	Clean Reads Q30	Clean Reads Ratio
N1307A	23.92	23.87	1.19	97.93	93.83	99.78
N1311A	23.92	23.66	1.18	97.95	93.95	98.91
N1314A	23.92	23.68	1.18	98.05	94.17	98.97
N1319A	23.92	23.88	1.19	97.99	93.99	99.82
N1322A	20.23	20.18	1.01	97.99	93.99	99.75
N1331A	20.15	19.9	0.99	97.9	93.79	98.75
N1341A	23.92	23.43	1.17	97.95	93.97	97.92
N1354A	23.92	22.31	1.12	97.9	94.06	93.27
N200A	23.92	23.72	10.19	97.84	93.61	99.14
N201A	23.92	23.74	10.19	97.91	93.76	99.24
N202A	22.18	22.12	10.11	97.94	93.8	99.74
N203A	23.92	23.73	10.19	97.83	93.58	99.2
N204A	23.92	23.7	10.18	97.91	93.86	99.05
N205A	23.92	23.77	10.19	97.8	93.46	99.34
N206A	23.92	23.64	10.18	98.15	94.55	98.82
N207A	21.74	21.68	10.08	97.88	93.66	99.74
N208A	26.1	24.49	10.22	98.19	94.85	93.83
N209A	23.92	23.43	10.17	98.14	94.52	97.93
N210A	23.92	23.86	10.19	97.92	93.68	99.72

Table S2 *continued*. The quality statistics of each sample.

Sample Name	Total Raw Reads	Total Clean Reads	Total Clean Bases	Clean Reads Q20	Clean Reads Q30	Clean Reads Ratio
N211A	22.42	21.73	10.09	98.34	95.01	96.93
N213A	23.92	23.62	10.18	98.11	94.48	98.75
N214A	23.92	23.43	10.17	98.2	94.73	97.96
N215A	23.92	23.75	10.19	97.92	93.76	99.29
N216A	23.92	23.47	10.17	98.17	94.62	98.09
N217A	23.92	23.84	10.19	98	93.95	99.67
N218A	23.92	22.72	10.14	98.1	94.49	94.96
N219A	23.92	23.82	10.19	98.1	94.26	99.56
N631A	23.92	23.77	10.19	98.2	94.59	99.38
N633A	23.92	23.73	10.19	97.81	93.5	99.17
N634A	23.92	23.45	10.17	97.78	93.45	98.03
N635A	23.92	23.52	10.18	98.09	94.36	98.31
N637A	23.92	23.69	10.18	97.76	93.32	99.02
N638A	23.92	23.87	10.19	98.27	94.76	99.76
N640A	23.92	23.86	10.19	98.15	94.49	99.74
N641A	23.92	23.86	10.19	98.07	94.3	99.73
N642A	23.92	23.86	10.19	98.11	94.37	99.76
N649A	23.92	23.75	10.19	97.86	93.62	99.29
N651A	23.92	23.87	10.19	98.11	94.3	99.76
N654A	23.92	23.82	10.19	97.63	92.99	99.56
T1000A	23.92	23.87	1.19	97.89	93.72	99.76
T1005A	26.1	23.06	1.15	97.92	94.13	88.36
T1007A	23.92	23.73	1.19	97.7	93.15	99.18
T1008A	23.92	23.26	1.16	97.76	93.48	97.21
T1010A	20.45	20.1	1.01	98.07	94.19	98.33
T1013A	26.1	23.48	1.17	97.93	94.17	89.98
T1016A	23.24	22.76	1.14	97.93	93.76	97.96
T1020A	23.14	22.78	1.14	97.93	93.74	98.47
T1023A	23.51	23.23	1.16	98	93.9	98.81
T1024A	23.92	23.72	1.19	98.01	93.94	99.14
T1026A	23.92	23.53	1.18	97.97	93.82	98.34
T1056A	23.92	23.78	1.19	97.92	93.63	99.41
T1058A	23.92	23.56	1.18	97.93	93.77	98.49
T1060A	21.57	20.64	1.03	97.78	93.59	95.7
T1061A	23.92	23.57	1.18	98.11	94.26	98.54
T1063A	23.92	23.78	1.19	97.96	93.75	99.4
T1067A	23.92	23.82	1.19	97.98	93.86	99.55
T1079A	23.92	23.84	1.19	97.96	93.79	99.67
T1082A	23.92	23.78	1.19	98.04	94.04	99.4
T1084A	23.52	23.08	1.15	98.02	94.08	98.1
T1085A	23.92	23.62	1.18	98.06	94.14	98.73
T1088A	15.91	15.21	0.76	98.22	94.86	95.63
T1089A	23.92	23.21	1.16	97.85	93.43	97.04
T1090A	26.1	25.37	1.27	97.94	94.02	97.2
T1091A	23.92	23.52	1.18	98.01	94.05	98.33

Table S2 *continued*. The quality statistics of each sample.

Sample Name	Total Raw Reads	Total Clean Reads	Total Clean Bases	Clean Reads Q20	Clean Reads Q30	Clean Reads Ratio
T1094A	23.85	23.15	1.16	97.97	94.06	97.08
T1096A	23.92	23.62	1.18	97.95	93.8	98.75
T1098A	23.92	23.77	1.19	97.65	92.74	99.38
T1104A	23.92	23.78	1.19	97.85	93.5	99.38
T1106A	23.92	23.72	1.19	97.57	92.65	99.16
T1108A	29.8	23.89	1.19	97.8	93.89	80.16
T1109A	23.92	23.66	1.18	97.7	93.23	98.89
T1112A	23.92	23.79	1.19	97.73	93.14	99.45
T1113A	23.92	23.62	1.18	97.91	93.71	98.73
T1114A	23.92	23.67	1.18	98.05	94.2	98.93
T1115A	23.92	23.78	1.19	97.91	93.71	99.4
T1119A	23.92	23.84	1.19	97.9	93.59	99.66
T1121A	23.92	23.78	1.19	97.66	92.89	99.38
T1123A	23.92	23.66	1.18	97.87	93.5	98.9
T1125A	23.52	22.98	1.15	98.13	93.34	97.69
T1127A	26.1	24.06	1.2	97.82	93.84	92.19
T1130A	23.92	23.73	1.19	97.86	93.51	99.18
T1136A	26.1	23.54	1.18	97.78	93.65	90.19
T1140A	23.92	23.67	1.18	97.83	93.44	98.93
T1142A	23.92	23.65	1.18	97.8	93.55	98.86
T1147A	32.53	21.03	1.05	97.98	94.17	64.66
T1148A	23.92	23.59	1.18	97.76	93.24	98.61
T1150A	23.92	23.71	1.19	97.84	93.42	99.13
T1152A	23.92	23.29	1.16	98.02	94.19	97.34
T1155A	23.92	23.63	1.18	97.94	93.8	98.75
T1156A	23.92	23.31	1.17	98.22	94.7	97.44
T1159A	23.92	23.48	1.17	97.99	93.9	98.14
T1161A	23.92	23.61	1.18	97.74	93.27	98.68
T1163A	22.5	19.63	0.98	97.88	94.02	87.26
T1164A	23.92	23.65	1.18	97.81	93.48	98.84
T1167A	23.92	23.37	1.17	97.87	93.73	97.68
T1168A	23.92	23.77	1.19	98.02	93.94	99.35
T1169A	23.92	23.24	1.16	97.95	94.18	97.16
T1172A	23.92	23.77	1.19	97.9	93.88	99.36
T1174A	23.92	23.68	1.18	97.94	94.13	98.98
T1179A	23.92	23.73	1.19	97.9	93.83	99.18
T1180A	21.34	21.09	1.05	97.9	93.86	98.82
T1181A	23.92	23.84	1.19	97.8	93.35	99.67
T1189A	23.22	23.06	1.15	97.94	93.92	99.32
T1190A	23.92	23.61	1.18	97.85	93.7	98.68
T1191A	23.92	23.72	1.19	97.76	93.42	99.15
T1193A	23.92	23.64	1.18	97.99	94.11	98.81
T1195A	23.91	23.54	1.18	97.91	93.89	98.45
T1199A	19.1	17.71	0.89	98.1	94.4	92.73
T1283A	23.92	23.43	1.17	98.02	94.07	97.96

Table S2 *continued*. The quality statistics of each sample.

Sample Name	Total Raw Reads	Total Clean Reads	Total Clean Bases	Clean Reads Q20	Clean Reads Q30	Clean Reads Ratio
T1284A	23.35	21.99	1.1	98.08	94.36	94.16
T1285A	23.92	23.5	1.18	98.18	94.57	98.25
T1286A	19.97	19.66	0.98	98.17	94.57	98.47
T1289A	23.92	23.46	1.17	98	93.95	98.07
T1291A	23.92	23.87	1.19	97.85	93.44	99.76
T1292A	23.92	23.39	1.17	97.96	93.94	97.79
T1295A	23.92	23.45	1.17	97.95	93.85	98.04
T1296A	23.92	23.39	1.17	98.11	94.32	97.77
T1297A	23.92	23.16	1.16	98.4	95.29	96.79
T1298A	23.92	22.82	1.14	97.99	94.08	95.39
T1299A	23.92	23.87	1.19	97.81	93.34	99.77
T1300A	23.92	23.45	1.17	98.06	94.26	98.03
T1301A	23.92	23.39	1.17	98.03	94.15	97.78
T1302A	23.92	23.49	1.17	98.1	94.31	98.19
T1303A	23.92	23.78	1.19	98.2	94.56	99.4
T1305A	23.92	23.44	1.17	98.16	94.54	97.98
T1306A	23.92	23.62	1.18	97.81	93.68	98.74
T1308A	23.92	23.53	1.18	97.8	93.71	98.36
T1309A	23.92	23.55	1.18	97.77	93.6	98.45
T1310A	23.92	23.54	1.18	98.03	94.41	98.41
T1313A	23.92	23.8	1.19	98.22	94.59	99.48
T1316A	21.72	21.32	1.07	98	94.22	98.13
T1317A	23.92	23.78	1.19	98.18	94.52	99.39
T1318A	23.92	23.76	1.19	98.18	94.51	99.3
T1321A	20.02	19.64	0.98	97.87	93.84	98.07
T1323A	23.92	23.79	1.19	98.17	94.43	99.42
T1324A	20.6	20.15	1.01	97.96	94.19	97.82
T1325A	23.92	23.58	1.18	97.94	94.1	98.56
T1326A	23.92	23.81	1.19	98.22	94.55	99.54
T1327A	23.92	23.55	1.18	97.84	93.75	98.42
T1328A	23.92	23.58	1.18	97.81	93.63	98.58
T1329A	23.92	23.53	1.18	97.88	93.91	98.36
T1330A	22.48	21.97	1.1	97.81	93.68	97.75
T1333A	19.05	18.54	0.93	98.02	94.42	97.28
T1335A	23.92	23.87	1.19	98.28	94.72	99.79
T1337A	22.3	21.89	1.09	97.95	94.11	98.17
T1338A	23.92	23.78	1.19	98.39	95.13	99.42
T1339A	23.92	23.52	1.18	97.79	93.71	98.33
T1340A	23.92	23.55	1.18	97.83	93.78	98.42
T1342A	23.92	23.87	1.19	98.16	94.4	99.78
T1343A	23.92	23.8	1.19	98.31	94.89	99.5
T1346A	23.92	23.78	1.19	98.26	94.74	99.39
T1347A	23.92	23.78	1.19	98.18	94.5	99.41
T1349A	23.92	23.7	1.18	97.68	93.12	99.06
T1352A	23.92	23.81	1.19	98.22	94.58	99.54

Table S2 *continued*. The quality statistics of each sample.

Sample Name	Total Raw Reads	Total Clean Reads	Total Clean Bases	Clean Reads Q20	Clean Reads Q30	Clean Reads Ratio
T1355A	23.92	23.78	1.19	98.14	94.38	99.42
T636A	23.92	23.62	10.18	97.87	93.5	98.73
T639A	30.45	24.07	10.2	97.9	94.02	79.06
T648A	23.92	23.68	10.18	97.98	93.92	98.97
T653A	22.43	21.83	10.09	98.09	94.24	97.33
T658A	23.92	23.65	10.18	98.03	94.01	98.84
T667A	23.92	23.64	10.18	97.87	93.57	98.81
T670A	23.92	23.62	10.18	97.83	93.45	98.73
T685A	21.74	21.46	10.07	98.11	94.26	98.73
T751A	23.92	23.67	10.18	97.87	93.5	98.96
T754A	23.92	23.64	10.18	97.82	93.33	98.82
T757A	23.92	23.65	10.18	97.98	93.92	98.85
T763A	23.92	23.65	10.18	98.03	94.02	98.86
T764A	23.92	23.63	10.18	98.01	93.9	98.79
T774A	23.83	23.48	10.17	98.16	94.43	98.53
T775A	23.92	23.66	10.18	97.71	93.17	98.9
T785A	23.75	23.16	10.16	98.23	93.63	97.52
T806A	23.92	23.61	10.18	97.79	93.45	98.69
T810A	23.92	22.91	10.15	97.71	93.39	95.78
T820A	23.92	23.58	10.18	97.66	93.06	98.57
T821A	23.92	23.63	10.18	98.15	94.35	98.79
T823A	22.46	21.13	10.06	97.99	94.18	94.08
T827A	23.92	23.59	10.18	98.18	94.45	98.62
T831A	23.92	23.53	10.18	97.79	93.48	98.35
T842A	23.92	23.58	10.18	97.8	93.48	98.58
T848A	23.92	23.53	10.18	97.81	93.36	98.36
T857A	23.92	23.62	10.18	98.25	94.78	98.71
T858A	23.92	23.22	10.16	97.88	93.83	97.08
T861A	23.92	23.38	10.17	97.59	92.91	97.74
T863A	23.92	23.65	10.18	98.1	94.23	98.88
T865A	23.92	23.58	10.18	97.95	93.82	98.57
T869A	23.92	23.29	10.16	97.88	93.78	97.36
T879A	23.92	22.44	10.12	97.69	93.42	93.79
T895A	23.92	23.63	10.18	97.97	93.84	98.77
T897A	23.92	23.63	10.18	97.79	93.45	98.79
T899A	23.92	23.62	10.18	97.7	93.2	98.75
T905A	23.92	23.6	10.18	97.86	93.54	98.66
T907A	26.1	24.05	10.2	97.72	93.47	92.16
T915A	23.92	23.64	10.18	97.97	93.9	98.8
T925A	20.87	20.16	10.01	97.85	93.88	96.6
T933A	23.92	23.74	10.19	97.66	92.99	99.24
T935A	18.99	14.52	0.739	7.809	3.747	6.45
T943A	23.92	23.62	10.18	98.07	94.18	98.75
T944A	13.81	10.9	0.5497	0.4392	0.6278	0.93
T948A	23.92	23.37	10.17	98.21	94.64	97.69

Table S2 *continued*. The quality statistics of each sample.

Sample Name	Total Raw Reads	Total Clean Reads	Total Clean Bases	Clean Reads Q20	Clean Reads Q30	Clean Reads Ratio
T951A	23.92	23.61	10.18	97.62	92.95	98.71
T953A	23.92	23.56	10.18	97.95	93.74	98.5
T956A	20.34	20.1	10.01	97.99	93.89	98.81
T957A	23.92	23.57	10.18	98.13	94.32	98.52
T959A	23.92	23.49	10.17	97.75	93.38	98.19
T960A	23.92	23.54	10.18	98	93.96	98.38
T961A	23.92	23.7	10.18	97.83	93.57	99.05
T967A	23.92	23.58	10.18	98.01	93.98	98.58
T972A	28.13	23.58	10.18	97.68	93.55	83.82
T974A	23.92	23.61	10.18	97.6	92.86	98.67
T977A	24.77	23.26	10.16	97.86	93.84	93.9
T984A	23.92	23.64	10.18	98.05	94.12	98.8
T986A	23.92	23.69	10.18	97.7	93.11	99.04
T997A	23.92	23.52	10.18	98.01	94.12	98.31

Table S3. Coefficients for genes in the risk model.

Gene	Coef
BMPER	0.151506596
SPSB4	0.171909418
SLAMF9	0.328271716
CLEC4E	-0.538517182
DKK1	0.001411042
IGSF1	0.345418968
RIMS2	1.098346519
SPP1	0.005621328
PTX3	0.070283848
CD276	0.379746233
BCHE	0.214988138
BMP2	0.089409309