



# A Four-Gene Signature Model Improves the Prediction of Distant Metastasis in Patients with Nasopharyngeal Carcinoma: A Retrospective, Three-Center Observational Study

Technology in Cancer Research & Treatment  
Volume 21: 1-11  
© The Author(s) 2022  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/15330338221080972  
journals.sagepub.com/home/tct  


Jinyuan Si, MD<sup>2</sup>, Xiuyong Ding, MD<sup>2</sup>, Zhuoxia Deng, MD<sup>1</sup>, Pu Li, MD<sup>2</sup>, Benjian Zhang, MD<sup>1</sup>, Guiping Lan, MD<sup>1</sup>, Bo Huang, MD<sup>1</sup>, Jinhui Liang, MD<sup>3</sup>, Zhenlin Wang, MD<sup>2,\*</sup>, and Yongfeng Si, MD<sup>1,\*</sup> 

## Abstract

**Background:** Similar to that in other malignant tumors, distant metastasis is one of the most important causes of poor prognosis in nasopharyngeal carcinoma (NPC). However, the genetic hallmarks and networks that regulate the distant metastasis of NPC are not fully understood. **Methods:** In this study, we performed high-throughput screening of mRNA expression profiles in 92 NPC samples collected from 3 hospitals and detected the mRNA expression levels of 31,503 genes in these samples. Gene functional enrichment analyses were performed using gene set enrichment analysis (GSEA). Least absolute shrinkage and selection operator (LASSO) was applied to select prognostic genes and a Cox proportional hazards regression model including these genes was constructed to predict prognosis. The Kaplan–Meier curve and time-dependent receiver operating characteristic (ROC) curve were plotted to assess the performance of this model. Univariate and multivariate analyses were performed using the Cox proportion hazard model to test the independence of prognostic effect of gene model and other clinical features. **Results:** A total of 1837 differentially expressed genes between patients with and without distant metastasis were identified in the training cohort, including 869 upregulated genes and 968 downregulated genes. Six gene sets, including the Wnt/ $\beta$  catenin signaling pathway, hedgehog (Hh) signaling pathway, Notch signaling pathway, mitotic spindle, apical surface, and estrogen response late, were enriched in patients with distant metastasis. A four-gene signature model was constructed in the training cohort, and according to the time-dependent ROC curve, this model had certain accuracy in predicting distant metastasis-free survival (DMFS) in both the training and validation cohorts. **Conclusion:** We developed a four-gene signature model that can evaluate the distant metastasis risk of NPC patients and may also provide novel therapeutic targets for NPC treatment in the near future.

## Keywords

prognostic factors, prediction, progression, oncology, outcomes

<sup>1</sup> The People's Hospital of Guangxi Zhuang Autonomous Region, Nanning, PR China

<sup>2</sup> Xuan Wu Hospital, Capital Medical University, Beijing, PR China

<sup>3</sup> Wuzhou Red Cross Hospital, Wuzhou, PR China

\*Senior authors contributed equally to this work.

## Corresponding Authors:

Yongfeng Si, MD, Department of Otorhinolaryngology – Head and Neck Oncology, The People's Hospital of Guangxi Zhuang Autonomous Region, 6 Taoyuan Street, Qingxiu District, Nanning 530021, PR China.  
Email: syfklxf@126.com

Zhenlin Wang, MD, Department of Otolaryngology – Head and Neck Surgery, Xuan Wu Hospital, Capital Medical University, 45 Changchun Street, Xicheng District, Beijing 100053, PR China.  
Email: wzl1812@163.com



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

## Abbreviations

AUC, area under the ROC curve; CIs, confidence intervals; CT, computed tomography; DFMS, distant metastasis-free survival; EB, Epstein–Barr; ECT, emission computed tomography; EMT, epithelial–mesenchymal transition; DEGs, differentially expressed genes; FPKM, fragments per kilobase million; GSEA, gene set enrichment analysis; Hh, hedgehog; HRs, hazard ratios; IMRT, intensity-modulated radiotherapy; LASSO, the least absolute shrinkage and selection operator; MRI, magnetic resonance imaging; NBI, narrow-band imaging; NPC, nasopharyngeal carcinoma; PET, positron emission tomography; ROC, receiver operating characteristic curve; RNA-Seq, RNA sequencing; SCUBE3, signal peptide-CUB-EGF-like domain-containing protein 3; TGF- $\beta$ , transforming growth factor  $\beta$ ; TPM, transcripts per kilobase million; UICC, Union for International Cancer Control/

## Introduction

Nasopharyngeal carcinoma (NPC) is a malignant tumor in the head and neck that is prevalent in South China, especially in Guangdong Province and the Guangxi Zhuang Autonomous Region.<sup>1</sup> In the past few decades, although substantial improvements have been made in radiotherapy and chemotherapy,<sup>2,3</sup> the clinical outcomes of NPC patients are still not satisfactory. Distant metastasis is the leading cause of treatment failure in NPC patients, and approximately 30% to 40% of locoregionally advanced NPC patients eventually develop distant metastasis after radical treatment.<sup>4</sup> At present, the main therapeutic strategy used to reduce the risk of distant metastasis in clinical practice is to treat patients at high risk of distant metastasis with systemic chemotherapy.<sup>5–7</sup> Although the risk of NPC patients developing distant metastasis can be partly characterized by TNM staging, the prediction accuracy is only 63.7%.<sup>7</sup> For instance, distant metastasis still occurs in 13% of stage N1 patients with low risk, but 65% of stage N3 patients with high risk do not develop distant metastasis.<sup>8</sup> Therefore, TNM staging is not precise enough to predict the risk of distant metastasis after radical therapy, and a more accurate method is warranted to determine the trend of distant metastasis after treatment and improve treatment strategies.

It is well known that the distant metastasis of cancer is an extremely complicated multistep process that consists of a succession of cell biological changes termed the invasion–metastasis cascade.<sup>9,10</sup> The mechanism underlying these biological changes is the high instability of the cancer genome (eg base deletions and insertions, gene amplifications, epigenetic dysregulation, and posttranscriptional and posttranslational modifications). Because mRNA expression profiles may be determined by multiple changes above, we performed high-throughput screening of mRNA profiles in 92 tissue samples of NPC with the goal of screening out the differentially expressed genes (DEGs) between NPC patients with and without distant metastasis. Subsequently, gene functional enrichment analysis was performed to elucidate and visualize the molecular mechanisms underlying the distant metastasis process of NPC. Then, through feature selection performed using least absolute shrinkage and selection operator (LASSO)-penalized Cox regression, four candidate genes were selected. Based on the regression coefficients of these 4 genes, a risk prediction model was established. This model was tested by performing survival analysis in high-risk and low-risk samples from both the training and validation cohorts, and the performance of the model was evaluated using the time-dependent receiver

operator characteristic (ROC) curve. To verify independence of prognostic effect of gene model and other clinical features, we further integrate patients of 2 validation cohorts (Beijing validation cohort and Wuzhou validation cohort) to perform univariate and multivariate analysis using Cox proportion hazard model.

Comparison of mRNA expression profiles between cancer cells and normal control cells or between cancer cells at different developmental stages or disease states could promote the discovery of characteristic gene signatures associated with oncogenesis and cancer progression. Recently, it has been suggested that certain gene expression patterns can be used as molecular models for early diagnosis, subgroup classification and even for risk prediction of distant metastasis in patients with locoregionally advanced NPC.<sup>11–14</sup> To this end, we performed this study with the aim of exploring the potential molecular mechanisms underlying the distant metastasis process of NPC and establishing a 4-gene model that can evaluate the risk of developing distant metastases and is suitable for patients with all stages of NPC.

## Materials and Methods

### Patient Selection and Tissue Sample Collection

A total of 92 patients were newly diagnosed with NPC between September 2007 and September 2012. Forty-four of the 92 patients were from People's Hospital of Guangxi Zhuang Autonomous Region, 25 of the 92 patients were from Xuanwu Hospital of Capital Medical University, and 23 of the 92 patients were from Wuzhou Red Cross Hospital. The present study was approved by the ethical committee of People's Hospital of Guangxi Zhuang Autonomous Region (No: KY-2019-001), Xuanwu Hospital (CLE2021-153), and Wuzhou Red Cross Hospital (S2019-35). All patients signed informed consent documents prior to participating in this study. The identity details of all patients enrolled in this study have been de-identified. The eligibility criteria of patients for inclusion in the study were as follows: (1) pathological confirmation of undifferentiated nonkeratinized carcinoma of the nasopharynx; (2) Union for International Cancer Control (UICC) staging system 2010 clinical classification of I to IVb, without a history of other malignant tumors or anticancer therapy; and (3) Karnofsky performance score  $\geq 70$ . The exclusion criteria included a history of severe systemic diseases, pregnancy or lactation, and the presence of a contradiction for receiving chemotherapy, radiotherapy, or surgery. Fresh NPC tissue samples of 92 NPC patients were obtained by nasopharynx biopsy under narrow-band imaging (NBI) endoscopy prior to

anticancer therapy and then frozen with liquid nitrogen ( $-196^{\circ}\text{C}$ ) until analysis.

### Pretreatment Evaluation of Patients

All of the patients underwent a pretreatment evaluation that included a precise clinical examination of the head and neck region, fiber optic nasopharyngoscopy, head and neck magnetic resonance imaging (MRI), chest x-ray, ultrasonography of the abdominal region, bone scan, and a complete blood count and biochemical profile.

### Patient Treatment

All patients were treated with intensity-modulated radiotherapy (IMRT) once a day 5 times a week. The target area was delineated based on the tumor boundary shown in MRI and computed tomography (CT). The prescribed doses were 69.76 to 76.3 Gy for PGTVnx, 64 to 70 Gy for PGTVnd, 59.4 to 64.0 Gy for PTV1, and 50.0 to 54.0 Gy for PTV2. All patients were administered with concurrent chemotherapy including cisplatin ( $40\text{ mg/m}^2$ ) or carboplatin ( $500\text{ mg/m}^2$ ), and fluorouracil ( $2000\text{ mg/m}^2$ ) with intervals of 21 days between cycles.

### Patient Follow-up

All the patients who had completed the NPC treatment attended follow-up visits every 3 months for the first year, every 6 months for the second to the fourth years, and annually thereafter. The follow-up methods were as follows: (1) All patients received Epstein-Barr (EB) virus serological examination and nasopharyngeal endoscopic examination, and endoscopically positive patients underwent nasopharyngeal biopsy. (2) All patients were confirmed by nasopharyngeal skull base MRI, chest x-ray, and cervical lymph node and abdominal ultrasonic examinations. (3) Patients who had been diagnosed with abnormal bone metabolism by emission computed tomography (ECT) scans received positron emission tomography (PET)/CT scans. The survival time was calculated from the date of first treatment completion to either the date of an event of interest (death caused by NPC or the development of distant metastasis) or the end of the follow-up.

### Procedure

Tissue RNA was isolated from fresh frozen tissue samples using the QIAGEN AllPrep DNA/RNA Mini Kit (Qiagen, Cat# 80204) according to the manufacturer's protocol. RNA degradation and contamination were monitored on 1% agarose gels. RNA purity was checked using a NanoPhotometer<sup>®</sup> spectrophotometer (Implen) or Nanodrop 2000 (Thermo Scientific). RNA concentration was measured using the Qubit<sup>®</sup> RNA Assay Kit on a Qubit<sup>®</sup> 2.0 Fluorometer (Life Technologies). RNA integrity was assessed using the RNA Nano 6000 Assay Kit on an Agilent Bioanalyzer 2100 system (Agilent Technologies).

For high-quality samples, a total amount of  $1.5\ \mu\text{g}$  RNA per sample was used as input material for the RNA sample

preparations. Sequencing libraries were generated using the NEBNext<sup>®</sup> UltraTM RNA Library Prep Kit for Illumina<sup>®</sup> (NEB, USA) following the manufacturer's recommendations, and index codes were added to attribute sequences to each sample. In brief, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations under elevated temperature in NEBNext First Strand Synthesis Reaction Buffer (5X). First-strand cDNA was synthesized using random hexamer primers and M-MuLV Reverse Transcriptase (RNaseH-). Second-strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. The remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation of the 3' ends of DNA fragments, NEBNext adaptors with hairpin loop structures were ligated to prepare for hybridization. To select cDNA fragments of correct length, the library fragments were purified with the AMPure XP system (Beckman Coulter). Then,  $3\ \mu\text{l}$  of USER Enzyme (NEB, USA) was used with size-selected, adaptor-ligated cDNA at  $37^{\circ}\text{C}$  for 15 min followed by 5 min at  $95^{\circ}\text{C}$  before PCR. PCR was performed with Phusion High-Fidelity DNA polymerase, universal PCR primers, and Index (X) Primer. Finally, the products were purified (AMPure XP system), and library quality was assessed on an Agilent BioAnalyzer 2100 system (Agilent Technologies). For low-quantity RNA samples (typically less than 100 ng of total RNA), the NEB Next rRNA Depletion Kit (Human/Mouse/Rat) (NEB E6310, USA) was used to deplete rRNA. Next, cDNA was synthesized from rRNA-depleted total RNA with the NEBNext RNA First-Strand Synthesis Module (NEB E7525S, USA), and double-stranded cDNA was generated from first-strand cDNA with the NEB Next mRNA Second-Strand Synthesis Module (NEB E6111S, USA) according to the manufacturer's instructions. DNA fragmentation and adapter ligation were carried out using the TruePrepTM DNA Library Prep Kit V2 for Illumina (Vazyme, China). The concentration of each library was quantified by the Qubit<sup>®</sup> RNA Assay Kit on a Qubit<sup>®</sup> 2.0 Fluorometer (Life Technologies) according to the manufacturer's instructions. The size distribution was evaluated using an Agilent high-sensitivity chip on an Agilent Bioanalyzer 2100 system (Agilent Technologies).

For mRNA sequencing, clustering of the index-coded samples was performed on a cBot Cluster Generation System using the HiSeq 4000 PE Cluster Kit (Illumina) or NextSeq 500/550 High Output Kit V2 (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina HiSeq 4000 or NextSeq 500 platform, and 150 bp paired-end reads were generated.

For data processing, Cutadapt was first used to trim adapter contamination and error-prone low-quality bases with the following command line parameters: `--quality-base = 33 --quality-cutoff = 20,20 --format = fastq -a CTGTCTCTTATACACATCT -A CTGTCTCTTATACACATCT -g AGATGTGTATAAGAGA CAG -G AGATGTGTATAAGAGACAG --times = 6 --minimum-length = 20 --max-n = 0.1 --trim-n`. Bowtie2 was used with default parameters to remove rRNA-containing reads by mapping reads to 5S rRNA, 5.8S rRNA, 18S rRNA, and 28S

rRNA sequences, as well as mitochondrial 16S rRNA and 12S rRNA sequences. High-quality reads were then aligned to the GRCh38 reference genome using STAR with the following parameters: `--outSAMUnmapped Within --outFilterType BySJout --outFilterMultimapNmax 20 --outFilterMismatch NoverLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --alignSJoverhangMin 8 --align SJDBoverhangMin 1 --sjdbScore 1 --runThreadN 2 --genome Load NoSharedMemory --outSAMtype BAM Unsorted --quantMode TranscriptomeSAM --readFilesCommand zcat`. Normalized gene expression (fragments per kilobase million [FPKM] and transcripts per kilobase million [TPM]) was calculated by RSEM with the following parameters: `-p 4 --seed-length 20 --paired-end --bam --estimate-rspd`.

### Statistical Analysis

**Calculation and justification of the sample size selected for this study.** To determine the sample size required for an RNA sequencing (RNA-Seq) experiment to identify differentially expressed genes (ie prognostic genes) between metastasis-positive group and metastasis-negative group, we used the R package `RnaSeqSampleSize` to calculate the minimum sample size. We assume that the minimum average read counts among the prognostic genes in the control group is 20, the dispersion for each gene is 0.1, and false discovery rate (FDR) threshold is set to be 0.2. We also suppose that the total number of genes for testing is 20,000 and the proportion of non-differentially expressed genes is 90%, and the minimum fold change is 1.5. By calculation, we will need to study 16 subjects in each group to be able to get an average power of 0.8.

**Identification of DEGs.** We select 16 pairs NPC tissues from 32 patients with or without distant metastasis after curative treatment. These 32 patients were selected from the total 92 patients enrolled in this study and were strictly matched by sex, age, TN stage, and therapeutic strategies to rule out the influence of these confounding factors on distant metastasis. Genes that were either not expressed or expressed at low levels (<5 reads in half of the samples) were removed prior to differential expression analysis. DESeq2 is a popular tool for the identification of DEGs. Since it requires raw counts for each gene as input, HTSeq was used to count reads mapping to each gene. The gene expression profiles of patients who developed metastasis within 5 years after diagnosis were compared to those of patients who remained metastasis-free for at least 5 years. Genes that met the following criteria were identified as DEGs: (1) fold change >1.5 or fold change < 0.667 and (2) adjusted *P* value <.2.

**Feature selection, risk model construction in the training cohort, and risk model evaluation in the validation cohort.** First, 48% (44 out of 92) of the RNA-Seq samples from People's Hospital of Guangxi Zhuang Autonomous Region were selected as the training cohort with the remaining samples from Xuanwu Hospital of Capital Medical University and Wuzhou Red

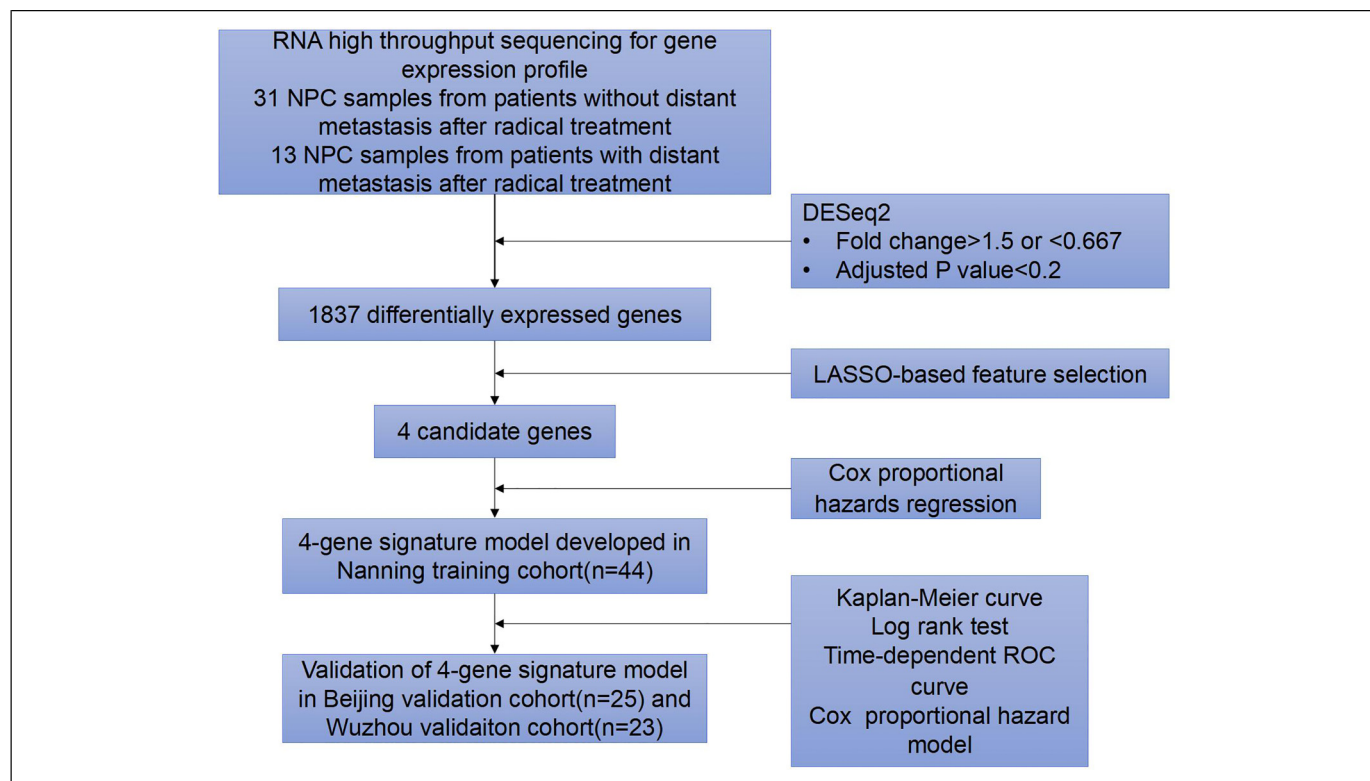
Cross Hospital were selected as the validation cohorts. Training samples were selected to ensure that the distribution of the selected samples was consistent with all samples in terms of age, sex, and clinical stage. After DEGs were identified from the training cohort, LASSO was applied to select prognostic genes. LASSO constructed a penalty function to obtain a more refined model so that it compressed some coefficients, and at the same time, it set some coefficients to zero. LASSO-based feature selection was conducted using the `glmnet` package in R software (version 3.4.0). The tuning parameter  $\lambda$  was determined according to the expected generalization error estimated from 10-fold cross-validation, and we adopted the  $\lambda$  value that gave minimum mean cross-validated error, known as  $\lambda_{\min}$ . A Cox proportional hazards regression risk model was then constructed to predict prognosis with genes selected by the LASSO algorithm. The risk scores of each sample from the training and validation cohorts were calculated according to the risk model. The respective medians of the 3 cohorts were used as the cut-off value to divide patients into high-risk and low-risk groups. Kaplan–Meier curves were plotted to compare the distant metastasis-free survival (DMFS) of high-risk and low-risk patients. *P* values and hazard ratios (HRs) with 95% confidence intervals (CIs) were generated by the log-rank test. The time-dependent ROC curve was plotted with the time ROC R package to characterize the discrimination potential of the risk score. The area under the ROC curve (AUC) was calculated as well. The univariate and multivariate analyses were performed using the Cox proportion hazard model.

**Gene set enrichment analysis.** Gene set enrichment analysis (GSEA) (version 3.0, <http://software.broadinstitute.org/gsea/downloads.jsp>) was performed between patients with and without metastasis within 5 years.<sup>14</sup> The Signal2Noise metric, which calculates the difference of means scaled by the standard deviation, was used to rank the genes. The ranked genes were compared against curated gene sets in the Molecular Signatures Database (<http://software.broadinstitute.org/gsea/msigdb>) for enrichment analysis. The `meandiv` normalization method was used to obtain the enrichment scores of the gene sets. A total of 1000 permutations were performed to generate a null distribution for the enrichment scores of the hallmark gene sets and functional annotation gene sets. The gene sets used for enrichment analysis included “`h.all.v6.2.symbols.gmt`” and “`c2.cp.kegg.v6.2.symbols.gmt`”. A cut-off of nominal  $P \leq .05$  was applied to identify significantly enriched gene sets.

## Results

### Patient Clinical Characteristics and Follow-up Outcome

latest patient follow-up visit occurred in July 2017. The follow-up time ranged from 2 to 114 months, with a median of 64 months. The flow chart of this study is shown in Figure 1. Among all 92 patients enrolled in this study, 44 patients from People's Hospital of Guangxi Zhuang



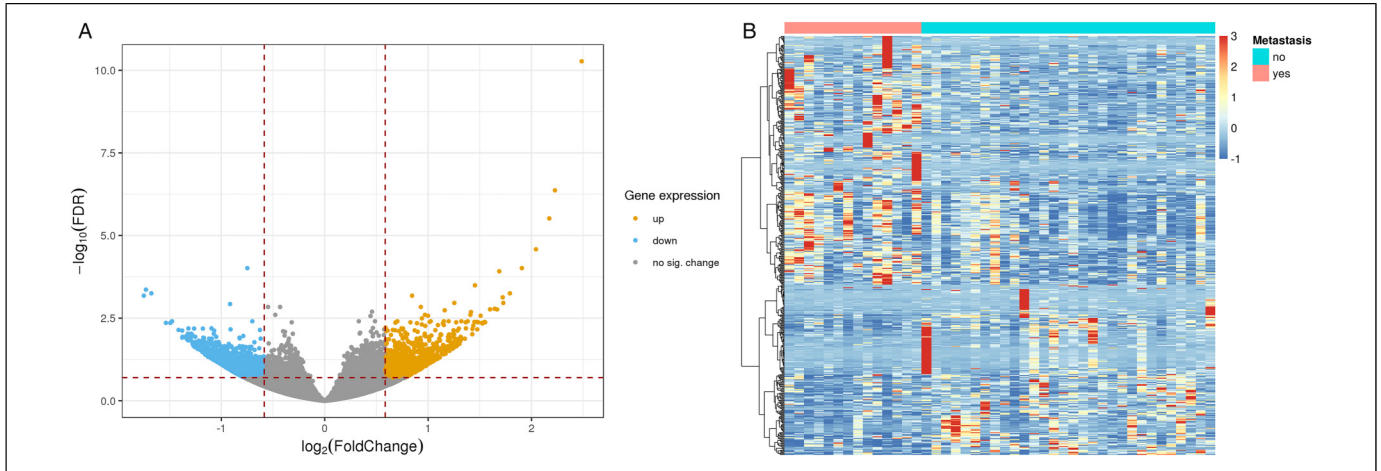
**Figure 1.** Flow chart of the study.

Abbreviations: LASSO: the least absolute shrinkage and selection operator; ROC curve: receiver operating characteristic curve.

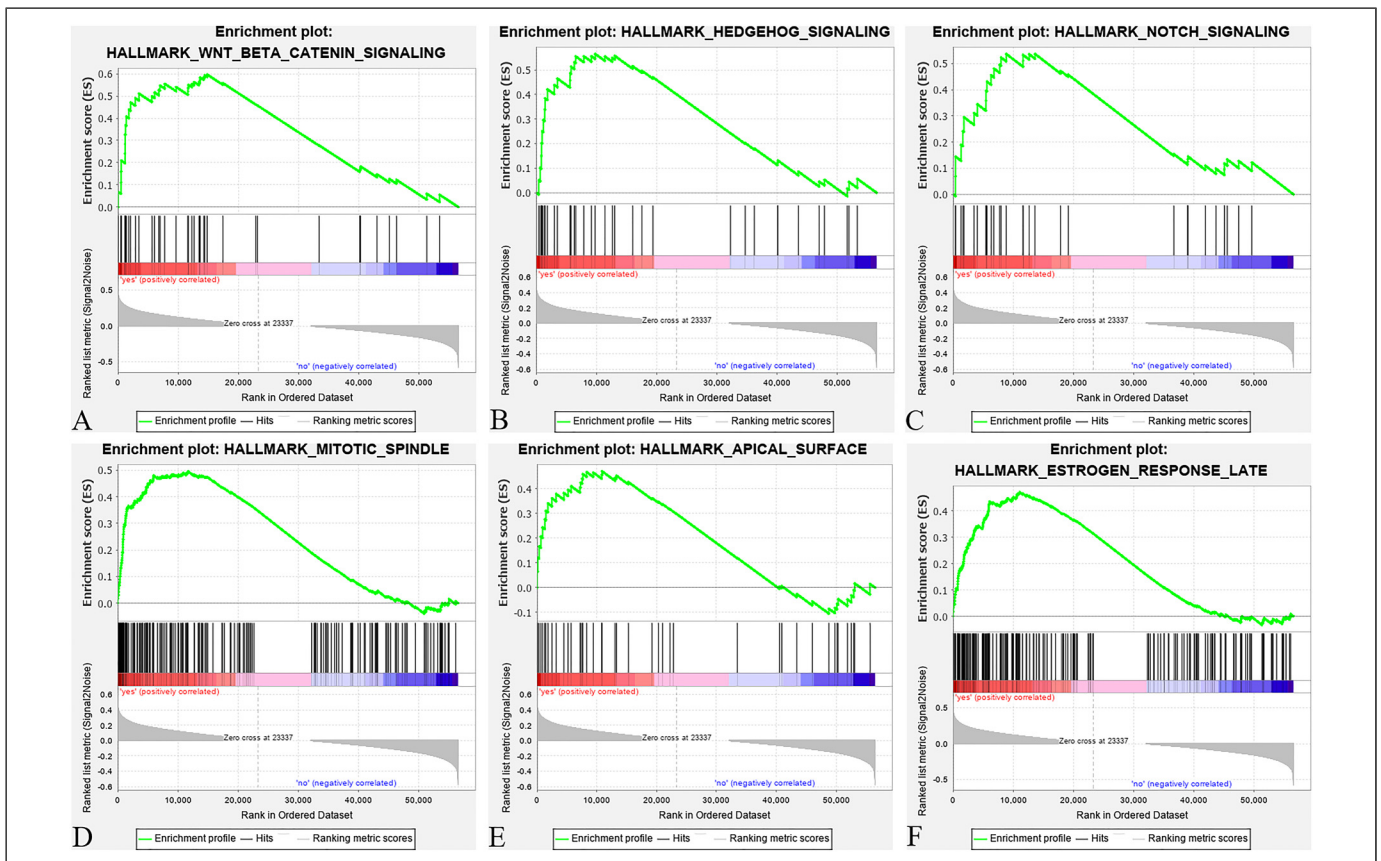
**Table 1.** The Clinical Characteristic of the 92 NPC Patients.

Variable	All NPC patients n = 92 (%)	Nanning training cohort n = 44 (%)	Beijing validation cohort n = 25 (%)	Wuzhou validation cohort n = 23 (%)
<b>Age (years)</b>				
≤45	38 (41.3)	19 (43.2)	12 (48.0)	7 (30.4)
>45	54 (58.7)	25 (56.8)	13 (52.0)	16 (69.6)
<b>Gender</b>				
<b>Male</b>	73 (79.3)	34 (77.3)	18 (72.0)	21 (91.3)
<b>Female</b>	19 (20.7)	10 (22.7)	7 (28.0)	2 (8.7)
<b>T stage</b>				
<b>T1</b>	3 (3.4)	3 (6.8)	0 (0.0)	0 (0.0)
<b>T2</b>	29 (31.5)	14 (31.8)	9 (36.0)	6 (26.1)
<b>T3</b>	27 (29.3)	11 (25.0)	7 (28.0)	9 (39.1)
<b>T4</b>	33 (35.8)	16 (36.4)	9 (36.0)	8 (34.8)
<b>N stage</b>				
<b>N0</b>	28 (30.4)	15 (34.1)	6 (24.0)	7 (30.4)
<b>N1</b>	28 (30.4)	16 (36.4)	6 (24.0)	6 (26.1)
<b>N2</b>	23 (25.0)	9 (20.5)	8 (32.0)	6 (26.1)
<b>N3</b>	13 (14.2)	4 (9.0)	5 (20.0)	4 (17.4)
<b>TNM stage</b>				
<b>I</b>	1 (1.1)	1 (2.3)	0 (0.0)	0 (0.00)
<b>II</b>	16 (17.4)	12 (27.3)	3 (12.0)	1 (4.3)
<b>III</b>	28 (30.4)	11 (25.0)	7 (28.0)	10 (43.5)
<b>IV</b>	47 (51.1)	20 (45.4)	15 (60.0)	12 (52.2)

Abbreviation: NPC: nasopharyngeal carcinoma.



**Figure 2.** Differentially expressed genes in patients with and without distant metastasis. (A) Volcano plot of 1837 differentially expressed genes that distinguish patients with and without distant metastasis. (B) Heatmap of 1837 differentially expressed genes that distinguish patients with and without distant metastasis. Patients are in columns and genes are in rows.

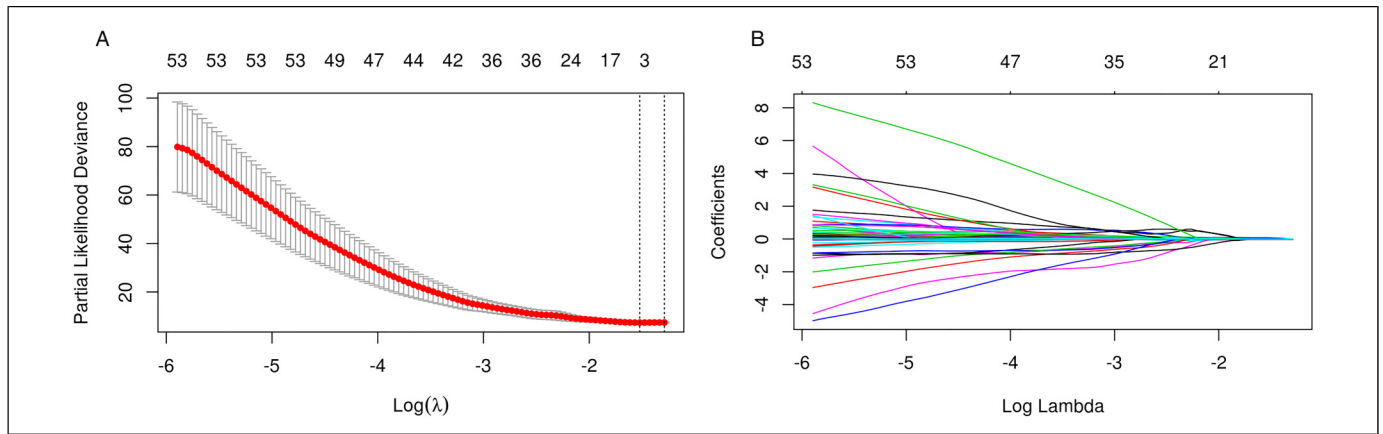


**Figure 3.** Gene set enrichment analysis (GSEA) of gene expression data in nasopharyngeal carcinoma (NPC). Six gene sets, including Wnt  $\beta$  catenin signaling pathway (A), hedgehog signaling pathway (B), Notch signaling pathway (C), mitotic spindle (D), apical surface (E), and estrogen response late (F), were enriched in the patients with distant metastasis.

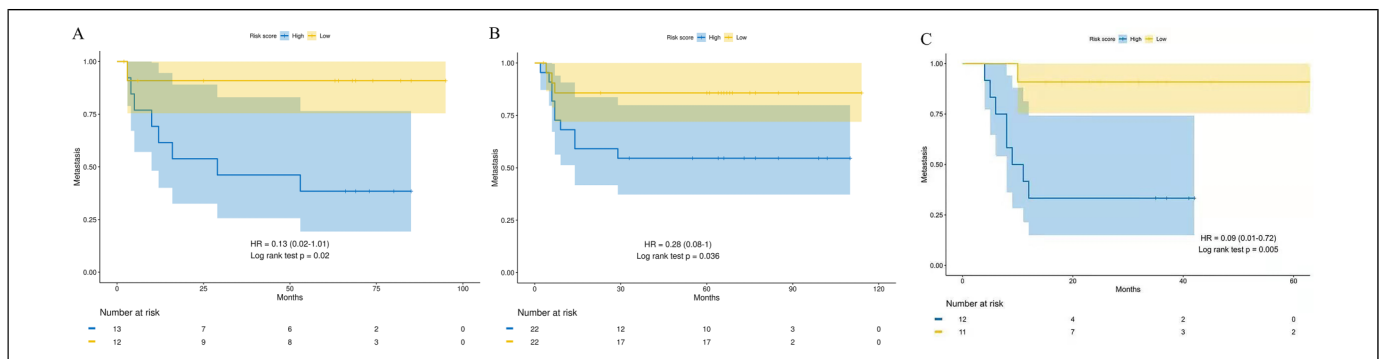
Autonomous Region were included in the training cohort, 25 patients from Xuanwu Hospital of Capital Medical University were included in the Beijing validation cohort, and 23 patients from Wuzhou Red Cross Hospital were included in the Wuzhou

validation cohort. All patients achieved clinical cure after the completion of the first treatment. Thirty-one of the 92 patients eventually developed distant metastasis, and 61 of the 92 patients did not develop distant metastasis by the end of the





**Figure 4.** The least absolute shrinkage and selection operator-based (LASSO-based) feature selection. (A) Tuning parameter ( $\lambda$ ) selection in the LASSO model used 10-fold cross-validation via minimum criteria. (B) LASSO coefficient profile plot produced against the log ( $\lambda$ ) sequence.



**Figure 5.** Kaplan–Meier curves of distant metastasis-free survival (DFMS) according to the four-gene signature model. (A) Nanning training cohort ( $n = 44$ ), (B) Beijing validation cohort ( $n = 25$ ), And (C) Wuzhou validation cohort ( $n = 23$ ).

follow-up. The clinical characteristics of the 92 patients are summarized in Table 1.

### DEG Screening and GSEA

In the bioinformatics analysis, 1837 mRNAs, including 869 upregulated genes and 968 downregulated genes, were found to be differentially expressed between 22 patients with nonmetastatic NPC and 47 patients with NPC who developed distant metastasis after treatment. The volcano plot and heatmap of DEGs are shown in Figure 2A and B. GSEA under the cut-off criteria defined by nominal  $P$  value  $< .05$  identified 6 gene sets, including the Wnt/ $\beta$  catenin signaling pathway, hedgehog (Hh) signaling pathway, Notch signaling pathway, mitotic spindle, apical surface, and estrogen response late, which were enriched in patients with distant metastasis (Figure 3).

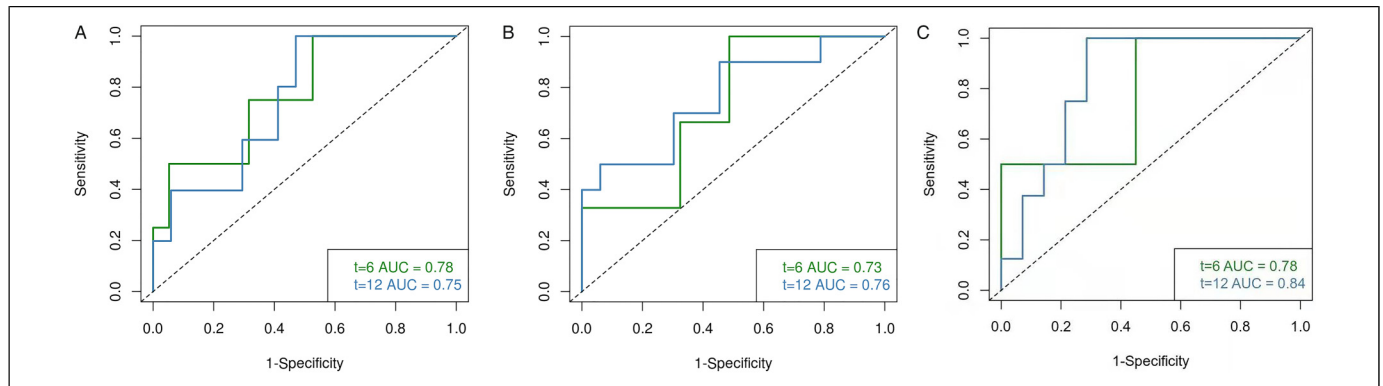
### Construction of a 4-Gene Signature Model in the Training Cohort

Through feature selection performed by LASSO Cox regression, a four-gene signature model was identified based on the optimal value of lambda (Figure 4A and B). The risk score

was calculated for each patient using the following formula derived from the expression level of these four genes weighted by their corresponding regression coefficient: Risk score =  $(0.0693 \times \text{expression of } LZTS2) + (0.2521 \times \text{expression of } SCUBE3) + (0.1988 \times \text{expression of } MAPK8IP2) - (0.5015 \times \text{expression of } FOXO6)$ . The patients in the training cohort were divided into a high-risk group ( $n = 22$ ) and a low-risk group ( $n = 22$ ) according to the median cut-off value. The Kaplan–Meier curve indicated that patients in the high-risk group had significantly worse DMFS than their low-risk counterparts (Figure 5A). The predictive performance of this four-gene signature model for DMFS was evaluated by a time-dependent ROC curve, and the AUC reached 0.78 at 6 months and 0.75 at 1 year (Figure 6A).

### Validation of the Four-Gene Signature Model in the Validation Cohort

To test the robustness of the four-gene signature model constructed from the training cohort, the patients in 2 validation cohorts were also stratified into high- and low-risk groups by the median cut-off value calculated with the same formula derived from the training cohort. Likewise, the patients in the



**Figure 6.** Time-dependent ROC curves of four-gene signature model for predicting the risk of distant metastasis. (A) Nanning training cohort, (B) Beijing validation cohort, and (C) Wuzhou validation cohort. Abbreviations: ROC curve: receiver operating characteristic curve.

high-risk group were more likely to suffer from distant metastasis and had a shorter survival time than those in the low-risk group (Figure 5B, and C). Moreover, the AUC of the four-gene signature model in Beijing validation cohort was 0.76 at 6 months and 0.73 at 1 year (Figure 6B), and was 0.78 at 6 months and 0.84 at 1 year when it in Wuzhou validation cohort (Figure 6C). To verify the independence of prognostic effect of gene model and other clinical features, we integrated the samples of 2 validation cohorts and performed univariate and multivariate analysis using Cox proportional hazard model. As summarized in Table 2 and Table 3, the four-gene signature model was an independent predictor of DMFS.

## Discussion

High-throughput RNA-Seq is a sensitive and reliable technique that can detect aberrant genetic expression in cancer patients.<sup>15</sup> Although a few studies have explored and elucidated the DEGs or differentially methylated markers between noncancerous nasopharyngeal epithelium and NPC or between different NPC subtypes,<sup>16,17</sup> the mRNA profiling differences between patients with different prognoses are still far from being understood. A recent study screened 137 DEGs between metastatic and nonmetastatic locoregionally advanced NPC samples and constructed a distant metastasis gene signature that consisted of 13 genes to predict the risk of distant metastasis in patients with locoregionally advanced NPC.<sup>14</sup> In the current study, a total of 1837 genes related to distant metastasis were identified, including 869 genes upregulated and 968 genes downregulated in patients with distant metastasis. Based on the identified DEGs, a novel prognostic model that can predict the risk of distant metastasis for patients with all stages of NPC was constructed in the Guangxi training cohort. A subsequent study showed that the novel prognostic model developed in the training cohort could categorize patients into high-risk and low-risk groups with significantly different metastasis-free survival rates. Furthermore, the AUC of this model was 0.78 and 0.75 at 6 months and 1 year of DMFS in the training cohort. Likewise, the AUC was 0.73 and 0.76 at 6 months and 1 year

of DMFS in Beijing validation cohort, and was 0.78 at 6 months and 0.84 at 1 year when it in Wuzhou validation cohort. The multivariate analysis shows that the four-gene signature model was an independent predictor of DMFS.

The prognostic model proposed in this study was composed of 4 candidate genes (eg *SCUBE3*, *LZTS2*, *MAPK8IP2*, and *FOXO6*). Signal peptide-CUB-EGF-like domain-containing protein 3 (*SCUBE3*) is a secreted cell-surface glycoprotein involved in promoting epithelial–mesenchymal transition (EMT), angiogenesis, and metastasis in lung cancer by activating the transforming growth factor  $\beta$  (TGF- $\beta$ ) receptor-Smad2/3 pathway.<sup>18</sup> Overexpression of *SCUBE3* results in the activation of TGF- $\beta$ 1, which can regulate the expression of *TWIST1* and thus contribute to the EMT and distant metastasis of breast cancer cells.<sup>19</sup> To our knowledge, there is currently no related study focusing on the correlation of *SCUBE3* and invasion or metastasis in NPC. Herein, our study suggests the possibility that the expression of *SCUBE3* may be associated with the distant metastasis of NPC for the first time. *MAPK8IP2* encodes a scaffold protein termed mitogen-activated protein kinase 8 interacting protein 2, which has been shown to interact with and regulate the activity of the *MAPK8/JNK1* and *MAP2K7/MKK7* kinases. In terms of cancer progression, the combined expression of *MAPK8IP2* and its interacting proteins (*FGF12* and *MAPK13*) has been found to correlate with poor survival in patients with esophageal squamous cell carcinoma.<sup>20</sup> *FOXO6* is a member of the Forkhead transcription factor family (*FOXO* family) and encodes a transcription factor named *FOXO6*. *FOXO6* has previously been identified to regulate memory consolidation and synaptic function.<sup>21</sup> Moreover, several studies found that *FOXO6* was involved in the regulation of very low-density lipoprotein production by promoting gluconeogenesis and integrating insulin signaling with microsomal triglyceride transfer protein.<sup>22,23</sup> However, the molecular function of *FOXO6* in the oncogenesis and progression of various cancers remains controversial. Hu et al.<sup>24</sup> reported that overexpression of *FOXO6* can inhibit the proliferation of lung cancer cells by inducing p53 accumulation, which indicates the tumor suppressor role of *FOXO6*. Conversely, Li



**Table 2.** Univariate Analysis with Cox Proportional Hazard Model for DMFS of NPC Patients.

Variables	P value	Hazard ratio	95% CI for hazard ratio	
			Lower	Upper
<b>Age</b>	.2161	1.00	0.95	1.06
<b>Gender</b>	.3419	1.12	0.31	4.06
<b>T stage</b>	.2828	1.47	0.80	2.70
<b>N stage</b>	.8683	1.29	0.77	2.16
<b>TNM stage</b>	.8725	1.43	0.75	2.73
<b>Gene model</b>	.0007	2.72	1.53	4.84

Abbreviations: DMFS, distant metastasis-free survival; NPC, nasopharyngeal carcinoma.

**Table 3.** Multivariate Analysis with Cox Proportional Hazard Model for DMFS of NPC Patients.

Variables	P value	Hazard ratio	95% CI for hazard ratio	
			Lower	Upper
<b>Age</b>	.7547	1.66	0.38	7.34
<b>Gender</b>	.5020	0.50	0.03	7.81
<b>T stage</b>	.3714	3.11	0.26	37.48
<b>N stage</b>	.2512	1.98	0.62	6.39
<b>TNM stage</b>	0.6239	1.98	0.62	6.39
<b>Gene model</b>	0.0008	3.29	1.64	6.59

Abbreviations: DMFS, distant metastasis-free survival; NPC, nasopharyngeal carcinoma.

et al.<sup>25</sup> found that FOXO6 was upregulated in gastric cancer and that overexpression of FOXO6 promoted cell proliferation by inducing C-myc expression. Therefore, the molecular roles of FOXO6 may be cell- or tissue-specific. The *LZTS2* gene is located on human chromosome 10q24.3, and its gene product has previously been reported to inhibit the cell growth and proliferation of NPC cell lines in vitro and in vivo, which suggests that *LZTS2* may function as a tumor suppressor in NPC.<sup>26,27</sup> However, the effects of *LZTS2* expression on the invasion and migration abilities of NPC cell lines were not elucidated in these studies. It is worth noting that according to the regression coefficient of *LZTS2* in our present study, patients with higher expression levels of *LZTS2* are more vulnerable to developing distant metastasis. In consideration of the contradictory outcomes above, we speculated that the role of *LZTS2* may vary during different periods of NPC progression, and the same phenomenon has been confirmed in TGF $\beta$ .<sup>28</sup> During the early stages of various cancers, TGF- $\beta$  has been reported to inhibit cancer cell proliferation by inducing the synthesis of CDKIs, p21 protein and 4EBP1 and suppressing the expression of the *MYC* gene.<sup>29–32</sup> In regard to distant metastasis, the binding of TGF- $\beta$  to its receptor TGF $\beta$ 2R induces the phosphorylation of the PAR6 protein and the degradation of the RhoA protein, thereby disrupting the intercellular junction between

cancer cells, accelerating the process of EMT and eventually promoting the development of distant metastasis.<sup>33,34</sup>

To gain more biological insight into the gene networks underlying distant metastasis in NPC, we performed GSEA and discovered that 6 gene sets associated with the Wnt/ $\beta$ -catenin signaling pathway, Hh signaling pathway, Notch signaling pathway, mitotic spindle, apical surface, and estrogen response late were enriched in NPC patients with distant metastasis. Of note, the Wnt/ $\beta$ -catenin signaling pathway has been reported to be involved in EMT and radioresistance in various cancers.<sup>35,36</sup> A recent study indicated that the positive expression of 2 Wnt/ $\beta$ -catenin signaling pathway-related proteins ( $\beta$ -catenin and c-MYC) was negatively linked with the survival rate of NPC patients.<sup>37</sup> Moreover, activation of the Wnt/ $\beta$ -catenin signaling pathway leads to NPC cell radioresistance and metastasis through nuclear translocation of  $\beta$ -catenin and transcriptional upregulation of HR pathway-related and EMT-related gene expression.<sup>38</sup> The Hh signaling pathway was first identified in the fruit fly and exerts its biological effects through a signaling cascade that culminates in a change in the balance between the activator and repressor forms of glioma-associated oncogene (Gli) transcription factors.<sup>39</sup> According to recent reports, the Hh signaling pathway is involved in the development of at least one-third of all malignant tumors<sup>40</sup> and therefore provides new insight into the development of molecular targets and tumor prevention strategies associated with the Hh signaling pathway. For instance, when surgery and radiotherapy are not effective treatment modalities, targeted Hh signaling pathway inhibition has been regarded as the treatment for locally aggressive basal cell carcinoma and metastatic basal cell carcinoma.<sup>41</sup>

Our study had some limitations. First, the sequencing data from 2 validation cohorts contained only 25 and 23 samples, respectively, and the current findings still need to be confirmed with a larger sample size. Second, the biological mechanisms by which the 4 candidate genes included in this prognostic model contribute to NPC metastasis remain unknown, and further assessment of their biological functions might provide novel targets for NPC treatment.

## Conclusion

Our study identified 1837 DEGs and 6 gene sets that may be involved in the distant metastasis process of NPC, and we further developed a novel four-gene signature model that can assess the risk of distant metastasis in NPC patients. This model may be used as a new tool to predict the prognosis of NPC patients.

## Acknowledgments

The authors would like to thank Huimin Wen, Yu Luo, Diange Li, and Huanxi Li for performing the sample preparation and RNA-Seq experiment. We also thank Jinsheng Tao for performing the statistical analysis.

## Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


## Funding

The authors received funding through Guangxi Zhuang Autonomous Region Key Research and Development Program (Grant No. AB19259002).

## Ethical Statement

The present study was approved by the ethical committee of People's Hospital of Guangxi Zhuang Autonomous Region (No: KY-2019-001), Xuanwu Hospital (CLE2021-153), and Wuzhou Red Cross Hospital (S2019-35). All patients provided written informed consent prior to enrollment in the study.

## ORCID iD

Yongfeng Si  <https://orcid.org/0000-0002-2143-4516>

## References

- Wei KR, Zheng RS, Zhang SW, Liang ZH, Li ZM, Chen WQ. Nasopharyngeal carcinoma incidence and mortality in China, 2013. *Chin J Cancer*. 2017;36(1):90. doi:10.1186/s40880-017-0257-9
- Lee AW, Lau WH, Tung SY, et al. Preliminary results of a randomized study on therapeutic gain by concurrent chemotherapy for regionally-advanced nasopharyngeal carcinoma: NPC-9901 trial by the Hong Kong nasopharyngeal cancer study group. *J Clin Oncol*. 2005;23(28):6966-6975. doi:10.1200/jco.2004.00.7542
- Lai SZ, Li WF, Chen L, et al. How does intensity-modulated radiotherapy versus conventional two-dimensional radiotherapy influence the treatment results in nasopharyngeal carcinoma patients? *Int J Radiat Oncol Biol Phys*. 2011;80(3):661-668. doi:10.1016/j.ijrobp.2010.03.024
- Hui EP, Leung SF, Au JS, et al. Lung metastasis alone in nasopharyngeal carcinoma: a relatively favorable prognostic group. A study by the Hong Kong nasopharyngeal carcinoma study group. *Cancer*. 2004;101(2):300-306. doi:10.1002/cncr.20358
- Al-Sarraf M, LeBlanc M, Giri PG, et al. Chemoradiotherapy versus radiotherapy in patients with advanced nasopharyngeal cancer: phase III randomized intergroup study 0099. *J Clin Oncol*. 1998;16(4):1310-1317. doi:10.1200/jco.1998.16.4.1310
- Lin JC, Jan JS, Hsu CY, Liang WM, Jiang RS, Wang WY. Phase III study of concurrent chemoradiotherapy versus radiotherapy alone for advanced nasopharyngeal carcinoma: positive effect on overall and progression-free survival. *J Clin Oncol*. 2003;21(4):631-637. doi:10.1200/jco.2003.06.158
- Wee J, Tan EH, Tai BC, et al. Randomized trial of radiotherapy versus concurrent chemoradiotherapy followed by adjuvant chemotherapy in patients with American joint committee on cancer/international union against cancer stage III and IV nasopharyngeal cancer of the endemic variety. *J Clin Oncol*. 2005;23(27):6730-6738. doi:10.1200/jco.2005.16.790
- Zhou GQ, Tang LL, Mao YP, et al. Baseline serum lactate dehydrogenase levels for patients treated with intensity-modulated radiotherapy for nasopharyngeal carcinoma: a predictor of poor prognosis and subsequent liver metastasis. *Int J Radiat Oncol Biol Phys*. 2012;82(3):e359-e365. doi:10.1016/j.ijrobp.2011.06.1967
- Talmadge JE, Fidler IJ. AACR Centennial series: the biology of cancer metastasis: historical perspective. *Cancer Res*. 2010;70(14):5649-5669. doi:10.1158/0008-5472.Can-10-1040
- Fidler IJ. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. *Nat Rev Cancer*. 2003;3(6):453-458. doi:10.1038/nrc1098
- Shi W, Bastianutto C, Li A, et al. Multiple dysregulated pathways in nasopharyngeal carcinoma revealed by gene expression profiling. *Int J Cancer*. 2006;119(10):2467-2475. doi:10.1002/ijc.22107
- Zeng Z, Zhou Y, Xiong W, et al. Analysis of gene expression identifies candidate molecular markers in nasopharyngeal carcinoma using microdissection and cDNA microarray. *J Cancer Res Clin Oncol*. 2007;133(2):71-81. doi:10.1007/s00432-006-0136-2
- Wang S, Li X, Li ZG, et al. Gene expression profile changes and possible molecular subtypes in differentiated-type nonkeratinizing nasopharyngeal carcinoma. *Int J Cancer*. 2011;128(4):753-762. doi:10.1002/ijc.25392
- Tang XR, Li YQ, Liang SB, et al. Development and validation of a gene expression-based signature to predict distant metastasis in locoregionally advanced nasopharyngeal carcinoma: a retrospective, multicentre, cohort study. *Lancet Oncol*. 2018;19(3):382-393. doi:10.1016/s1470-2045(18)30080-9
- Teng CF, Li TC, Huang HY, et al. Next-generation sequencing-based quantitative detection of hepatitis B virus pre-S mutants in plasma predicts hepatocellular carcinoma recurrence. *Viruses*. 2020;12(8):796. doi:10.3390/v12080796
- Ali SM, Yao M, Yao J, et al. Comprehensive genomic profiling of different subtypes of nasopharyngeal carcinoma reveals similarities and differences to guide targeted therapy. *Cancer*. 2017;123(18):3628-3637. doi:10.1002/cncr.30781
- Hui L, Zhang J, Ding X, Guo X, Jang X. Identification of potentially critical differentially methylated genes in nasopharyngeal carcinoma: a comprehensive analysis of methylation profiling and gene expression profiling. *Oncol Lett*. 2017;14(6):7171-7178. doi:10.3892/ol.2017.7083
- Wu YY, Peck K, Chang YL, et al. SCUBE3 is an endogenous TGF- $\beta$  receptor ligand and regulates the epithelial-mesenchymal transition in lung cancer. *Oncogene*. 2011;30(34):3682-3693. doi:10.1038/onc.2011.85
- Yang X, Hu J, Shi C, Dai J. Activation of TGF- $\beta$ 1 pathway by SCUBE3 regulates TWIST1 expression and promotes breast cancer progression. *Cancer Biother Radiopharm*. 2020;35(2):120-128. doi:10.1089/cbr.2019.2990
- Bhushan A, Singh A, Kapur S, et al. Identification and validation of fibroblast growth factor 12 gene as a novel potential biomarker in esophageal cancer using cancer genomic datasets. *Omics*. 2017;21(10):616-631. doi:10.1089/omi.2017.0116
- Salih DA, Rashid AJ, Colas D, et al. Foxo6 regulates memory consolidation and synaptic function. *Genes Dev*. 2012;26(24):2780-2801. doi:10.1101/gad.208926.112
- Kim DH, Perdomo G, Zhang T, et al. Foxo6 integrates insulin signaling with gluconeogenesis in the liver. *Diabetes*. 2011;60(11):2763-2774. doi:10.2337/db11-0548

23. Kim DH, Zhang T, Lee S, et al. Foxo6 integrates insulin signaling with MTP for regulating VLDL production in the liver. *Endocrinology*. 2014;155(4):1255-1267. doi:10.1210/en.2013-1856
24. Hu HJ, Zhang LG, Wang ZH, Guo XX. Foxo6 inhibits cell proliferation in lung carcinoma through up-regulation of USP7. *Mol Med Rep*. 2015;12(1):575-580. doi:10.3892/mmr.2015.3362
25. Li Q, Cui L, Du Z-D, et al. FOXO6 Promotes gastric cancer cell tumorigenicity via upregulation of C-myc. *FEBS Lett*. 2013;587(14):2105-2111. doi:10.1016/j.febslet.2013.05.027
26. Cabeza-Arvelaiz Y, Thompson TC, Sepulveda JL, Chinault AC. LAPSER1: a novel candidate tumor suppressor gene from 10q24.3. *Oncogene*. 2001;20(46):6707-6717. doi:10.1038/sj.onc.1204866
27. Xu S, Li Y, Lu Y, et al. LZTS2 Inhibits PI3 K/AKT activation and radioresistance in nasopharyngeal carcinoma by interacting with p85. *Cancer Lett*. 2018;420(1):38-48. doi:10.1016/j.canlet.2018.01.067
28. Ikushima H, Miyazono K. TGFbeta signalling: a complex web in cancer progression. *Nat Rev Cancer*. 2010;10(6):415-424. doi:10.1038/nrc2853
29. Datto MB, Li Y, Panus JF, Howe DJ, Xiong Y, Wang XF. Transforming growth factor beta induces the cyclin-dependent kinase inhibitor p21 through a p53-independent mechanism. *Proc Natl Acad Sci U S A*. 1995;92(12):5545-5549. doi:10.1073/pnas.92.12.5545
30. Hannon GJ, Beach D. p15INK4B is a potential effector of TGF-beta-induced cell cycle arrest. *Nature*. 1994;371(6494):257-261. doi:10.1038/371257a0
31. Yagi K, Furuhashi M, Aoki H, et al. c-myc is a downstream target of the smad pathway. *J Biol Chem*. 2002;277(1):854-861. doi:10.1074/jbc.M104170200
32. Azar R, Alard A, Susini C, Bousquet C, Pyronnet S. 4E-BP1 Is a target of Smad4 essential for TGFbeta-mediated inhibition of cell proliferation. *Embo J*. 2009;28(22):3514-3522. doi:10.1038/emboj.2009.291
33. Hurd TW, Gao L, Roh MH, Macara IG, Margolis B. Direct interaction of two polarity complexes implicated in epithelial tight junction assembly. *Nat Cell Biol*. 2003;5(2):137-142. doi:10.1038/ncb923
34. Ozdamar B, Bose R, Barrios-Rodiles M, Wang HR, Zhang Y, Wrana JL. Regulation of the polarity protein Par6 by TGFbeta receptors controls epithelial cell plasticity. *Science*. 2005;307(5715):1603-1609. doi:10.1126/science.1105718
35. Ma X, Yan W, Dai Z, et al. Baicalein suppresses metastasis of breast cancer cells by inhibiting EMT via downregulation of SATB1 and Wnt/beta-catenin pathway. *Drug Des Devel Ther*. 2016;10(2):1419-1441. doi:10.2147/dddt.S102541
36. Zhao Y, Tao L, Yi J, Song H, Chen L. The role of canonical Wnt signaling in regulating radioresistance. *Cell Physiol Biochem*. 2018;48(2):419-432. doi:10.1159/000491774
37. Pang Q, Hu W, Zhang X, Pang M. Wnt/beta-catenin signaling pathway-related proteins (DKK-3, beta-catenin, and c-MYC) are involved in prognosis of nasopharyngeal carcinoma. *Cancer Biother Radiopharm*. 2019;34(7):436-443. doi:10.1089/cbr.2019.2771
38. Yang XZ, Chen XM, Zeng LS, et al. Rab1A promotes cancer metastasis and radioresistance through activating GSK-3beta/Wnt/beta-catenin signaling in nasopharyngeal carcinoma. *Aging (Albany NY)*. 2020;12(20):20380-20395. doi:10.18632/aging.103829
39. Skoda AM, Simovic D, Karin V, Kardum V, Vranic S, Serman L. The role of the hedgehog signaling pathway in cancer: a comprehensive review. *Bosn J Basic Med Sci*. 2018;18(1):8-20. doi:10.17305/bjbm.2018.2756
40. Murone M, Rosenthal A, de Sauvage FJ. Hedgehog signal transduction: from flies to vertebrates. *Exp Cell Res*. 1999;253(1):25-33. doi:10.1006/excr.1999.4676
41. Atwood SX, Li M, Lee A, Tang JY, Oro AE. GLI Activation by atypical protein kinase C  $\iota/\lambda$  regulates the growth of basal cell carcinomas. *Nature*. 2013;494(7438):484-488. doi:10.1038/nature11889