


Identification of Diagnostic and Prognostic Biomarkers in Nasopharyngeal Carcinoma Using Integrated Transcriptomics and Elastic Net Survival Analysis



Nur Aziz^{1,*} , Laily Rahmawati² and Jae Youl Cho³

¹Department of Histology and Cell Biology, Faculty of Medicine, Public Health, and Nursing, Universitas Gadjah Mada, Indonesia

²Department of Molecular Biology, Faculty of Medicine, Universitas Negeri Yogyakarta, Indonesia

³Department of Integrative Biotechnology, and Biomedical Institute for Convergence at SKKU, Sungkyunkwan University, Suwon16419, Republic of Korea

Abstract:

Introduction: Nasopharyngeal carcinoma (NPC) is a malignant tumor with distinct molecular features, underscoring the need for reliable biomarkers to improve diagnosis, prognosis, and therapeutic strategies.

Methods: We analyzed transcriptomic data from GEO datasets (GSE12452, GSE53819, and GSE102349) to identify diagnostic and prognostic biomarkers. Differential expression analysis was performed to detect potential markers, while survival analysis was conducted using Cox proportional hazards (Cox-PH) modeling and log-rank tests. Elastic Net regression was used to refine the gene signature. RNA-protein expression concordance was validated using the Cancer Cell Line Encyclopedia (CCLE) dataset.

Results: Differential expression analysis revealed 591 genes as potential diagnostic markers. Survival analysis identified 54 genes with dual diagnostic and prognostic relevance. Elastic Net regression refined this to an 11-gene signature, which stratified patients into high- and low-risk groups, significantly predicting progression-free survival (log-rank $p = 0.0035$). Five genes (*BUB1B*, *GAS2L3*, *NFE2L3*, *OIP5*, and *PDGFRL*) were identified as potential oncogenic drivers, while six (*CD1D*, *CYP4B1*, *IL33*, *KLF2*, *NAPSB*, and *VILL*) were implicated as tumor suppressors. Six genes (*BUB1B*, *GAS2L3*, *IL33*, *OIP5*, *PDGFRL*, and *VILL*) showed strong RNA-protein expression concordance in the CCLE dataset.

Discussion: This study reveals previously unreported cancer-associated genes (*NAPSB*, *GAS2L3*, *NFE2L3*, *PDGFRL*, *CD1D*, *CYP4B1*, *KLF2*) in NPC while validating established biomarkers (*BUB1B*, *OIP5*, *IL33*, *VILL*). Our findings expand NPC molecular characterization but require further clinical validation.

Conclusion: This study presents a robust gene signature for NPC, offering valuable insights into tumor progression and providing a foundation for advancing diagnostic strategies, improving prognostic stratification, and developing targeted therapies.

Keywords: Biomarkers, Diagnostic, Elastic Net, NPC, Prognostic.

© 2025 The Author(s). Published by Bentham Open.

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Public License (CC-BY 4.0), a copy of which is available at: <https://creativecommons.org/licenses/by/4.0/legalcode>. This license permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

*Address correspondence to this author at the Department of Histology and Cell Biology, Faculty of Medicine, Public Health, and Nursing, Universitas Gadjah Mada, Indonesia; nur.aziz20@ugm.ac.id

Cite as: Aziz N, Rahmawati L, Cho J. Identification of Diagnostic and Prognostic Biomarkers in Nasopharyngeal Carcinoma Using Integrated Transcriptomics and Elastic Net Survival Analysis. Open Bioinform J, 2025; 18: e18750362408821. <http://dx.doi.org/10.2174/0118750362408821250821062122>



Received: April 30, 2025
Revised: July 12, 2025
Accepted: July 16, 2025
Published: August 23, 2025



Send Orders for Reprints to
reprints@benthamscience.net

1. INTRODUCTION

NPC is a relatively rare malignancy; however, its incidence is disproportionately high in specific geographic regions, including Southeast Asia and North Africa. In Indonesia, NPC was the fourth most prevalent cancer among males in 2022, with a total of 14,497 newly diagnosed cases reported across all age groups [1]. Due to its nonspecific symptoms, NPC is often diagnosed at advanced stages, highlighting the critical need for molecular diagnostic tools to facilitate early detection, guide treatment decisions, monitor disease progression, and predict prognosis [2].

Biomarkers for NPC can be broadly categorized into two groups: *Epstein-Barr* virus (EBV)-related biomarkers and cellular biomarkers. EBV-related biomarkers, such as EBV DNA, *EBNA1*, *LMP1/2*, *EBER 1/2*, and miRNA *BART*, have been extensively studied and are widely used for NPC diagnosis and monitoring [2]. However, these biomarkers have limitations, as not all NPC patients exhibit detectable EBV reactivation [3, 4]. This highlights the need for complementary cellular biomarkers that reflect the intrinsic molecular characteristics of NPC.

Transcriptomic analysis has emerged as a powerful tool for biomarker discovery, enabling the identification of comprehensive gene expression patterns associated with tumorigenesis, progression, and therapeutic responses [5]. By leveraging high-throughput sequencing technologies, transcriptomics provides a global view of gene expression, uncovering molecular signatures that serve as diagnostic, prognostic, and predictive biomarkers. This approach allows for the identification of unique gene expression patterns specific to tumors, offering deeper insights into disease pathogenesis and potential therapeutic targets [6]. Recent studies have leveraged transcriptomic data to identify novel biomarkers for NPC. For instance, bioinformatics analysis of transcriptomic datasets has identified *RASGRP2*, *TTC9*, *CD37*, *DPM3*, and *ARHGAP4* as potential prognostic markers [7].

Unlike previous studies that primarily focused on prognostic biomarkers, this study aims to identify both diagnostic and prognostic biomarkers for NPC through integrated bioinformatics analysis of publicly available transcriptomic datasets. By focusing on cellular biomarkers, we seek to complement existing EBV-based approaches and provide a more comprehensive understanding of NPC biology. Our findings are expected to contribute to the development of improved diagnostic and prognostic tools, ultimately enhancing patient outcomes in NPC management.

2. MATERIALS AND METHODS

2.1. Study Design

This research is a quantitative, analytical observational study aimed at identifying diagnostic and prognostic biomarkers for nasopharyngeal carcinoma (NPC) using publicly available transcriptomic datasets. The study employed a secondary data analysis approach, using datasets obtained from the Gene Expression Omnibus (GEO), specifically GSE12452, GSE53819, and GSE102349.

2.2. Study Population and Data Collection

Datasets for diagnostic gene screening were selected based on the following criteria: (1) inclusion of both normal nasopharyngeal tissue and NPC tissue, and (2) a minimum of 10 samples per group. Exclusion criteria included: (1) datasets lacking raw expression data, or (2) datasets without clinical annotations or normal-tumor classification information. Sample sizes and clinical annotations varied across datasets and were obtained directly from the GEO database.

2.3. Variables and Measures

The primary variables were gene expression levels across samples. The main outcomes were differential gene expression (NPC vs. control) and progression-free survival, where survival data were available.

2.4. Dataset Acquisition and Preprocessing

Gene expression datasets for nasopharyngeal carcinoma (NPC) were obtained from the Gene Expression Omnibus (GEO) database (<https://www.ncbi.nlm.nih.gov/gds>) using the search criteria: "(nasopharyngeal carcinoma) AND 'Homo sapiens'[porgn: txid9606]". This search identified a total of 20 datasets. Each dataset was manually reviewed based on predefined inclusion criteria. The datasets screened for this study were provided in Table S1. Following this filtering process, two datasets (accession numbers: GSE12452, comprising 10 normal and 31 NPC tissue samples [8] and GSE53819, consisting of 18 normal and 18 NPC samples [9]) were selected for further analysis. Samples were classified into "Normal" and "Cancer" groups based on clinical annotations. Data were retrieved using the *GEOquery* package in R version 4.4.2 and *RStudio*. Raw expression matrices were normalized, log2-transformed, and filtered to remove missing values and transcripts without gene symbols. Principal Component Analysis (PCA) was performed using the *PCAtools* package [10] on log2-transformed expression matrices after filtering out the bottom 10% of low-variance genes, to assess global expression patterns and to evaluate whether normal and tumor samples exhibit distinct clustering based on transcriptomic profiles.

2.5. Identification of Differentially Expressed Genes (DEGs)

Differential expression analysis was performed using the *limma* package in R [11]. A linear model was fitted to the expression data, and empirical Bayes moderation was applied to compute log2 fold changes (logFC) and adjusted p-values. Genes with an absolute logFC > 1 and an adjusted p-value < 0.05 were considered statistically significant DEGs. Genes without valid gene symbols or duplicate entries were removed by retaining the entry with the smallest p-value per gene. The top 25 upregulated and top 25 downregulated DEGs were selected based on fold change magnitude. Expression values of these 50 genes were extracted from the normalized expression matrix, log-transformed, and converted to Z-scores by row-wise standardization. A

heatmap was generated using the ComplexHeatmap package in R [12], applying a diverging color scale for Z-scores and annotating columns based on sample groups (normal vs. cancer).

2.6. Survival Analysis

Among the 20 datasets previously filtered, only GSE102349 contained prognostic data. This dataset, comprising 113 NPC samples, was used to identify prognostic biomarkers based on RNA-seq expression profiles. For each gene, expression levels were dichotomized into “High” and “Low” groups using the median expression value as the cutoff. Survival analysis was performed using the Kaplan-Meier method and Cox proportional hazards (Cox-PH) models. The lower expression group was used as the reference. Genes with a p-value < 0.05 in the Cox-PH model were considered significant prognostic markers. Kaplan-Meier survival curves were generated for each gene, and hazard ratios (HR) with 95% confidence intervals (CI) were calculated.

2.7. Functional Annotation and Integration of Prognostic-Diagnostic Markers

Prognostic and diagnostic roles of genes were classified as “Oncogenic” or “Tumor Suppressor” based on two criteria: (1) expression trends in tumor versus normal tissues, and (2) hazard ratios (HR) from Cox proportional hazards (Cox-PH) models. Genes upregulated in tumor tissues were classified as “Oncogenic,” while those downregulated in tumor tissues were classified as “Tumor Suppressors.” Similarly, genes with HR > 1 in Cox-PH analysis were labeled as “Oncogenic,” whereas genes with HR < 1 were labeled as “Tumor Suppressors.” Genes demonstrating consistent roles as either oncogenic or tumor suppressors across both diagnostic (DEGs) and prognostic (Cox-PH) analyses were categorized as “Consistent Prognostic-Diagnostic Markers.”

2.8. Elastic Net Model Analysis

The Elastic Net model, implemented using the *glmnet* package in R, was applied to refine the 54 candidate genes identified from CoxPH analysis. By combining L1 (LASSO) and L2 (Ridge) regularization, Elastic Net balances feature selection and multicollinearity handling, addressing limitations of CoxPH in high-dimensional data [13]. A survival object was created using “Time to event” and “Event” status, and the predictor matrix included expression values of the 54 candidate genes. Ten-fold cross-validation determined the optimal regularization parameter (lambda), and the final model identified genes with non-zero coefficients as significant. This approach ensured a robust, interpretable subset of survival-associated genes, reducing overfitting and improving generalizability. Similar machine learning approaches, including Elastic Net, have been successfully applied in other diseases such as Parkinson’s disease and breast cancer for robust biomarker selection [14, 15], supporting its use in our study.

2.9. RNA and Protein Correlation in Cancer Cell Lines Database

Gene and protein expression data were obtained from the CCLE database using the *depmap* R package [16], specifically version 22Q2 [17] with a snapshot date of 2024-10-24. RNA-seq expression data, measured in Transcripts Per Million (TPM), were retrieved using the *depmap_TPM()* function, while protein expression levels were extracted using the *depmap_proteomic()* function. The datasets were also available for download from <https://depmap.org/portal>. The datasets were filtered to remove missing values and grouped by tissue or lineage based on cell line annotations. To assess RNA-protein concordance, candidate gene and protein expression data were integrated by matching cell lines across the datasets. Pearson correlation analysis was performed to evaluate the relationship between RNA and protein expression levels, and the results were visualized using scatter plots with regression trend lines to illustrate the degree of correlation.

2.10. Cross-Dataset Validation of Diagnostic Gene Signatures

To validate the diagnostic potential of 11 candidate genes across independent datasets, we generated a comprehensive heatmap by integrating statistical test results from multiple GEO datasets. The datasets GSE12452 and GSE53819, previously used for diagnostic gene discovery, served as training sets, while GSE218847, GSE61218 [18], GSE40290, GSE34573 [19], GSE64634 [20], GSE227541 [21], GSE134886 [22], and GSE118719 [23] were used as validation datasets. Validation datasets were manually curated based on the inclusion criterion that the sample size ratio between cancer and normal groups did not exceed fivefold, to reduce sample imbalance bias. For each gene in each dataset, a color matrix was constructed: red indicated higher expression in Nasopharyngeal Carcinoma (NPC) samples, blue indicated higher expression in normal samples, and gray represented no significant differential expression or unavailable data (e.g., probe not present or other technical reasons). A separate matrix overlaid statistical significance with symbols. Given the small sample sizes in some datasets, these results were interpreted cautiously and primarily served as visual cross-validation of consistency across datasets.

3. RESULTS

3.1. Identification of Diagnostic Biomarkers

To identify diagnostic biomarkers distinguishing normal nasopharyngeal tissue from NPC, we analyzed microarray data from GSE12452 (10 normal, 31 NPC) and GSE53819 (18 normal, 18 NPC). Principal Component Analysis (PCA) was conducted using the *PCAtools* package to evaluate the global expression patterns and sample clustering within each GSE series. In GSE12452, the first two principal components (PC1 and PC2) explained 20.00% and 12.70% of the total variance, respectively. In GSE53819, PC1 and PC2 accounted for 19.01% and

12.53% of the variance (Fig. 1A). In both datasets, PCA revealed a clear separation between normal and cancer groups along the first principal component, indicating that the primary source of expression variability corresponds to the disease state. These findings support the biological relevance and internal consistency of each dataset, confirming their suitability for downstream differential expression analysis.

Differential expression analysis identified 1,332 DEGs in GSE12452 and 2,444 DEGs in GSE53819 (adjusted p-value < 0.05, $|\log_{2}FC| > 1$) (Fig. 1B). A heatmap of the top 25 upregulated and top 25 downregulated DEGs across the two diagnostic datasets revealed clear and consistent

transcriptomic separation between normal and cancer samples. These patterns highlight robust differential expression profiles associated with NPC (Fig. 1C). Notably, this study emphasizes identifying consistent DEG patterns for diagnostic biomarker discovery, rather than investigating the biological mechanisms of the top DEGs. To this end, a Venn diagram analysis was conducted, revealing 597 overlapping DEGs between the two datasets, with 591 genes showing consistent expression patterns (*i.e.*, the same direction of regulation in both datasets) (Fig. 1D). This consistency across independent datasets enhances the reliability of these genes as potential diagnostic biomarkers for NPC.

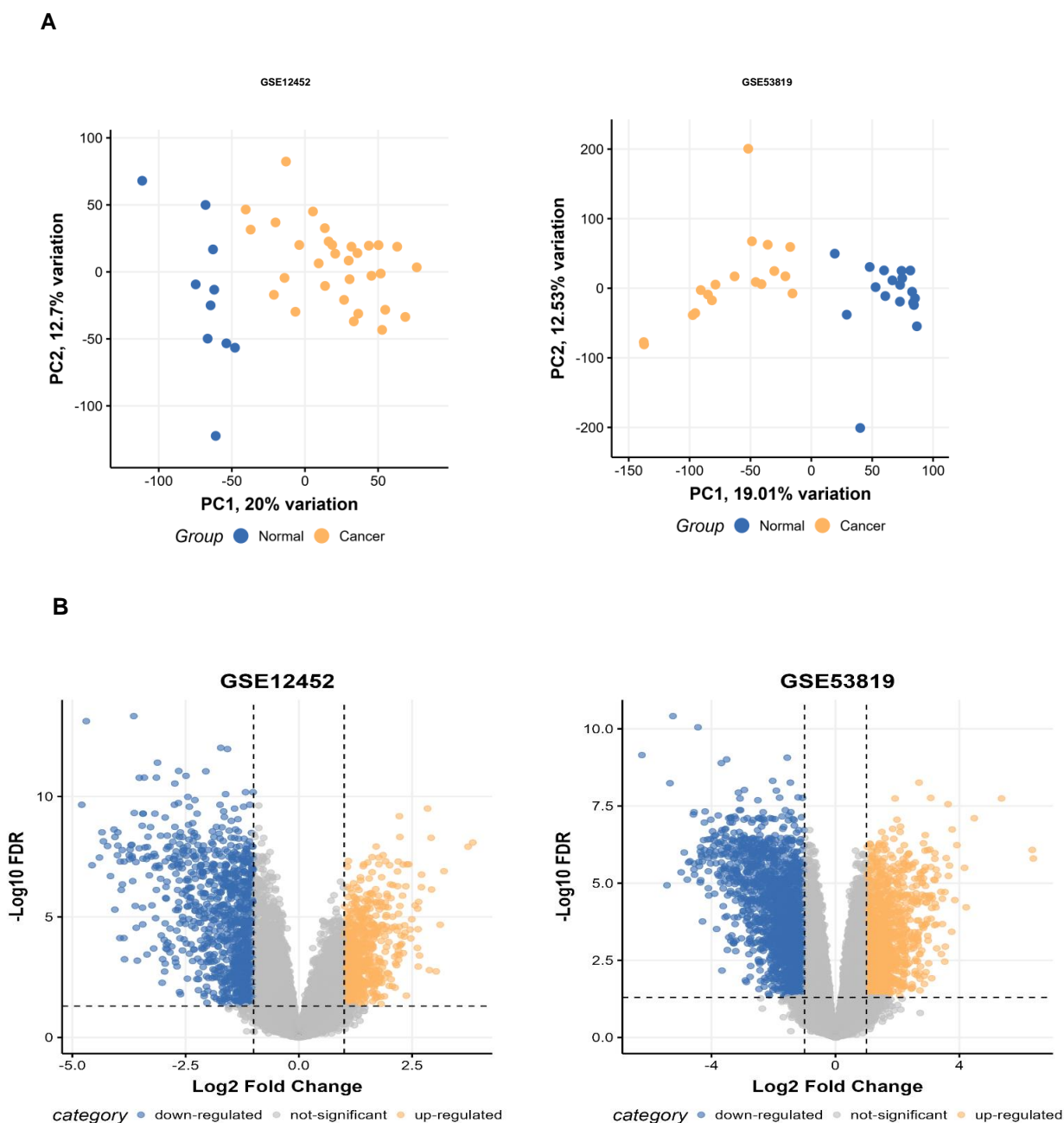


Fig. 1 contd.....

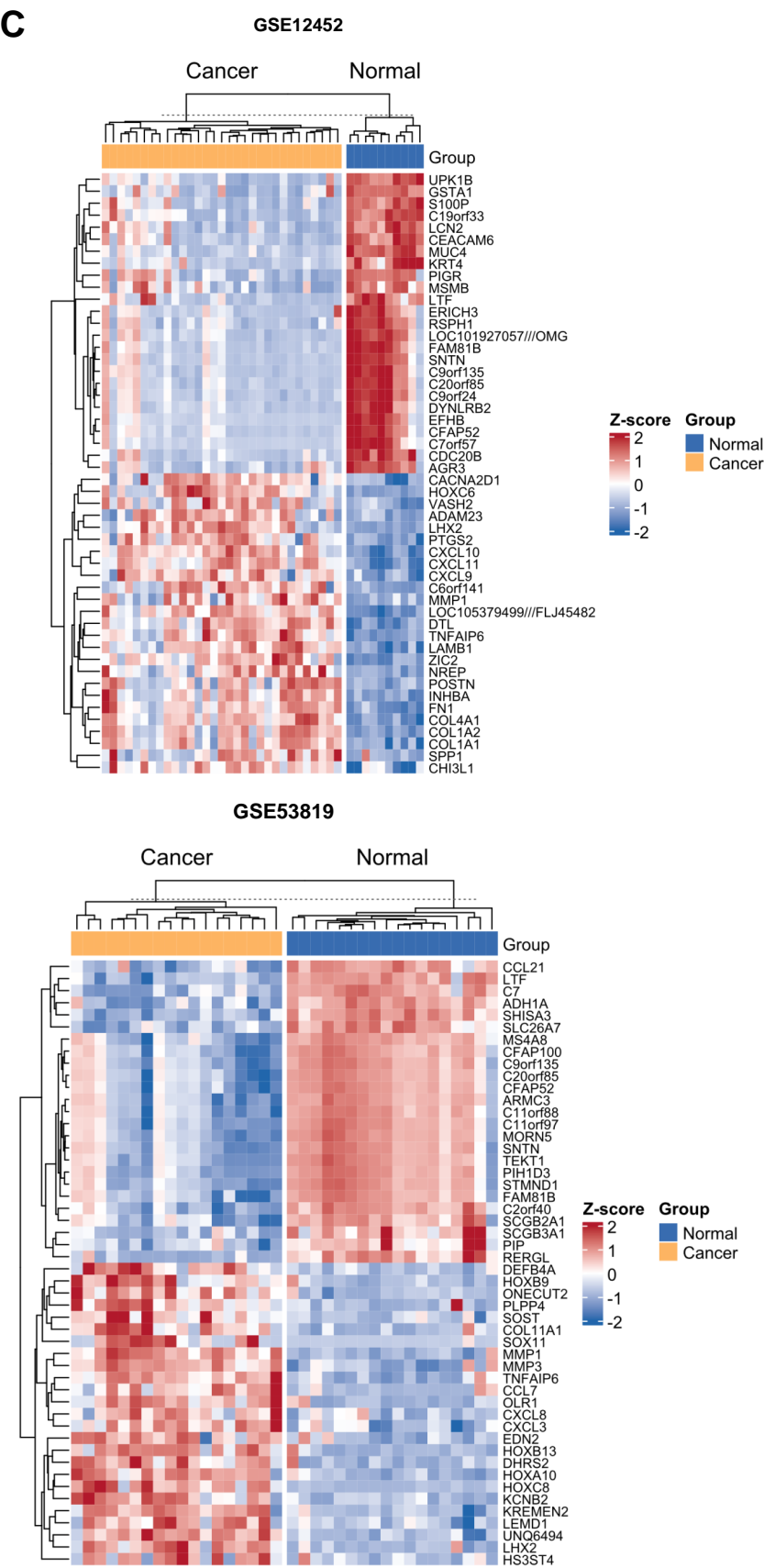


Fig. 1 contd.....

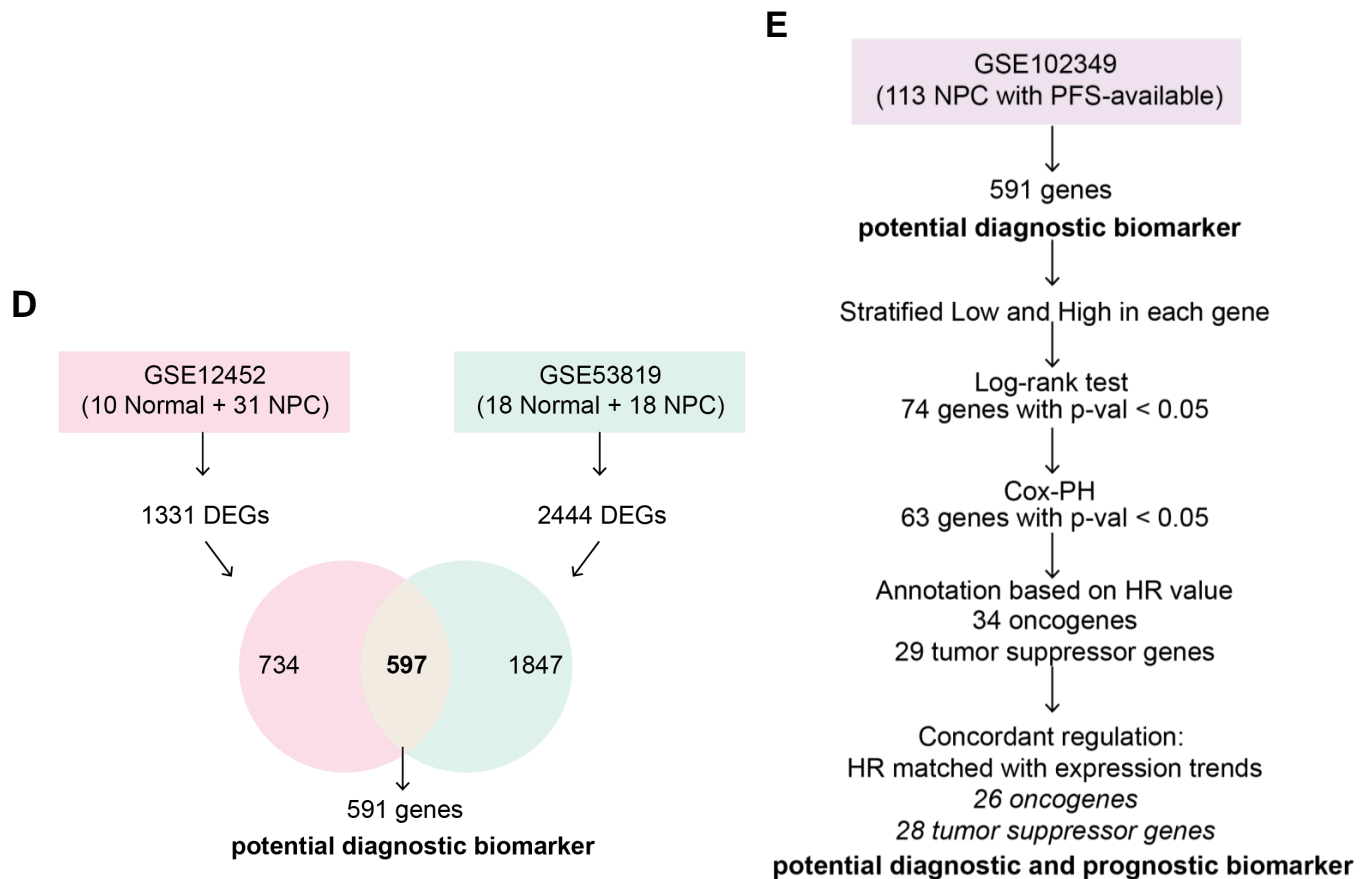


Fig. (1). Identification of potential diagnostic and prognostic biomarkers for NPC. (A) Principal component analysis highlights clear differences in gene expression between normal and cancer samples in the GSE12452 and GSE53819 datasets. (B) Volcano plot showing differentially expressed genes (DEGs) from nasopharyngeal carcinoma microarray datasets (GSE12452 and GSE53819). Yellow and blue dots represent significantly upregulated and downregulated genes, respectively ($|\log FC| > 1$, adjusted p-value < 0.05). (C) Heatmap showing the top 25 upregulated and top 25 downregulated differentially expressed genes (DEGs) in GSE12452 and GSE53819 (D) Analytical framework for identifying potential diagnostic biomarkers for NPC. (E) Analytical framework for identifying potential diagnostic and prognostic biomarkers for NPC.

3.2. Identification of Prognostic Biomarkers

To identify prognostic biomarkers associated with progression-free survival (PFS) in NPC, we performed survival analysis on RNA-seq data from the GSE102349 dataset. Gene expression levels were stratified into “High” and “Low” groups based on median expression values, and survival analysis was conducted using the log-rank test and Cox proportional hazards (Cox-PH) model. An analytical framework for identifying potential diagnostic and prognostic biomarkers was illustrated in Fig. (1D).

Among the 591 potential diagnostic biomarker genes, a total of 74 genes were significantly associated with PFS based on the log-rank test (p-value < 0.05). Of these, 63 genes also showed significant associations in the Cox proportional hazards (Cox-PH) model (p-value < 0.05), demonstrating their robust prognostic potential. These genes were further classified based on their hazard ratios (HR): genes with $HR > 1$ were categorized as “Oncogenic,” while those with $HR < 1$ were classified as “Tumor Suppressors.” Additionally, we evaluated the

concordant regulation by comparing HR with expression differences between normal and cancer tissues. This analysis revealed that 54 out of the 63 genes exhibited concordant regulation, reinforcing their dual diagnostic and prognostic significance.

Among the 54 candidate diagnostic-prognostic biomarkers, 26 genes were identified as oncogenic. These genes were upregulated in NPC compared to normal nasopharyngeal tissues and were associated with a higher risk of disease progression. Key oncogenic genes include *KIF14*, *NEK2*, *DTL*, *EXO1*, *SPP1*, *MAD2L1*, *CENPH*, *SEMA6A*, *TTK*, *NFE2L3*, *ANLN*, *PDGFRL*, *CDCA2*, *PAPPA*, *RCN1*, *GAS2L3*, *BUB1B*, *OIP5*, *KIF23*, *FANCI*, *PRC1*, *TOP2A*, *PSMC3IP*, *KIF18B*, *BRIP1*, and *BIRC5*.

Conversely, 28 genes were identified as tumor suppressors. These genes exhibited lower expression in NPC tissues compared to normal nasopharyngeal tissues, and higher expression of these genes was associated with a lower risk of disease progression. Key tumor suppressor genes include *CYP4B1*, *FCRL4*, *FCRL2*, *FCRL1*, *CD1D*,

FCRLA, FCMR, CR2, CR1, VILL, DTHD1, STAP1, ADH1B, BANK1, IL33, PAX5, PTGDS, MS4A1, KLRB1, SLC16A7, CD19, PLCG2, P2RX5, CD79B, KLF2, CD22, NAPSB, and VPREB3. This functional classification provides valuable insights into the biological roles of these prognostic biomarkers, enhancing our understanding of NPC progression and highlighting their potential as therapeutic targets.

3.3. Refinement of Candidate Genes Using Elastic Net

To enhance the selection of prognostic genes, the Elastic Net model was applied using the *glmnet* package in R, narrowing down the 54 candidate genes derived from the Cox proportional hazards (CoxPH) analysis. The Elastic Net analysis identified 11 genes with non-zero coefficients, indicating their potential association with survival outcomes. Cross-validation curve for the Elastic Net model was shown in Fig. (2A). Among these, 5 genes (e.g., *GAS2L3*, *OIP5*, *NFE2L3*, *PDGFRL*, *BUB1B*) exhibited positive coefficients, suggesting that higher expression levels are associated with an increased risk of disease progression. These genes were previously categorized as oncogenic. Conversely, 6 genes (e.g., *VILL*, *IL33*, *KLF2*, *NAPSB*, *CYP4B1*, *CD1D*) showed negative coefficients, implying that higher expression levels may be protective and associated with better survival outcomes. These genes

were previously categorized as tumor suppressors. The strongest positive association was observed for *GAS2L3* (coefficient = 0.47), while the strongest negative association was found for *VILL* (coefficient = -0.22). This refined subset of genes, derived from the initial 54 candidates, highlights their potential as biomarkers for survival prediction and demonstrates the utility of Elastic Net in handling high-dimensional data to identify robust, interpretable gene signatures.

Patients were stratified into “High Risk” and “Low Risk” groups based on the median risk score derived from the Elastic Net model. The risk score for each patient was calculated as the weighted sum of the expression values (in TPM) of the 11 selected genes, using the coefficients obtained from the final Elastic Net Cox model. The formula for the risk score is: Risk Score = $(-0.0176 \times CYP4B1) + (-0.0056 \times CD1D) + (-0.2185 \times VILL) + (0.0926 \times NFE2L3) + (0.0925 \times PDGFRL) + (-0.0393 \times IL33) + (0.4696 \times GAS2L3) + (0.0329 \times BUB1B) + (0.1264 \times OIP5) + (-0.016 \times KLF2) + (-0.0527 \times NAPSB)$. Kaplan-Meier survival analysis revealed a significant difference in progression-free survival between the two groups (log-rank p-value = 0.0035) (Fig. 2B). Patients in the “High Risk” group exhibited significantly lower progression-free survival compared to those in the “Low Risk” group, further validating the prognostic utility of the identified gene signature.

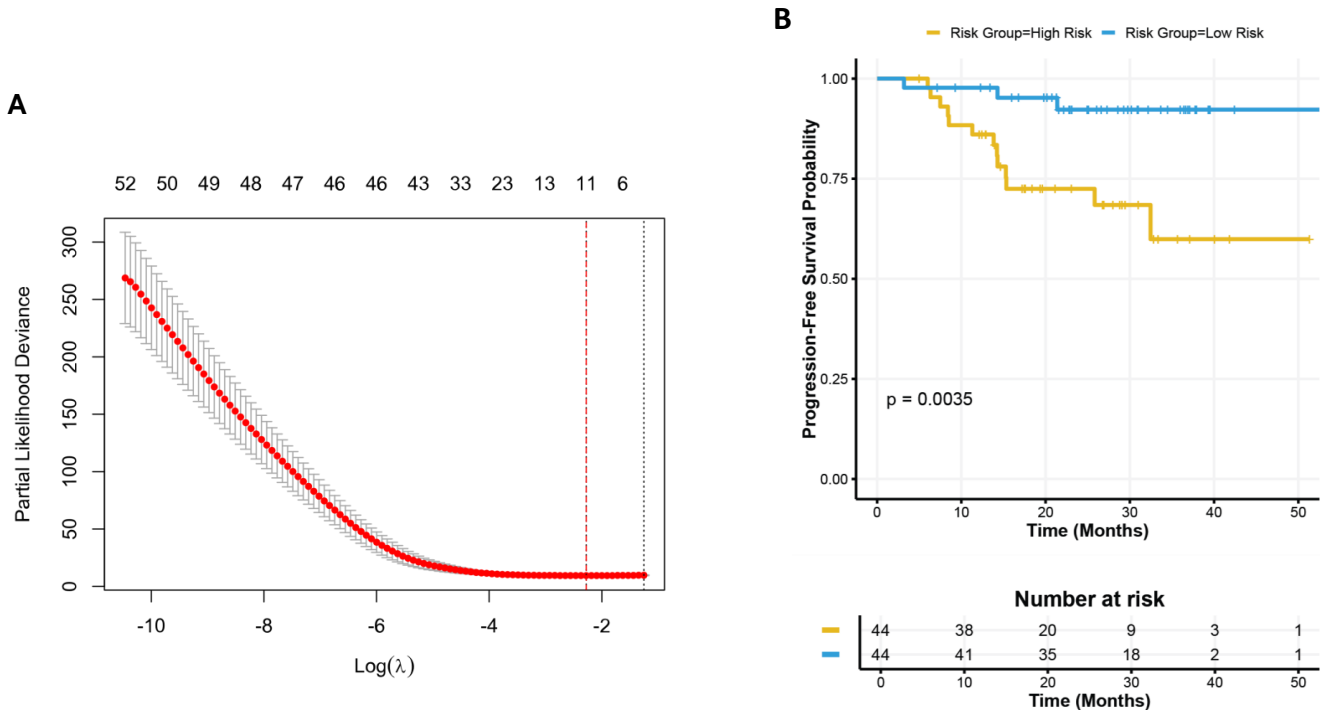


Fig. (2). Elastic net regression and survival analysis. (A) Cross-validation curve for the Elastic Net model, showing the relationship between log(lambda) and partial likelihood deviance. The vertical dashed line indicates the optimal lambda value (λ_{min}) selected based on the minimum cross-validation error. (B) Progression-free survival curves stratified by risk groups (“High Risk” and “Low Risk”) based on the median risk score derived from the Elastic Net model. The log-rank test p-value (p = 0.0035) indicates a significant difference in progression-free survival between the two groups. The risk table below the plot displays the number of patients at risk over time.

3.4. Exploration on the Potential 11 Candidate Diagnostic and Prognostic Genes

The expression profiles of the 11 candidate genes were compared between normal nasopharyngeal tissue and NPC tissue using the GSE12452 and GSE53819 datasets. All 11 genes were significantly differentially expressed between normal and NPC tissues ($p < 0.01$, Fig. 3A), suggesting their potential roles in NPC pathogenesis and their value as diagnostic markers. Kaplan-Meier survival analysis revealed that the expression levels of these genes were significantly associated with PFS in NPC patients (log-rank test, $p < 0.05$, Fig. 3B). Stratification of patients into high- and low-expression groups demonstrated

distinct survival outcomes, further supporting the prognostic relevance of these genes.

The forest plot (Fig. 3C) illustrates the hazard ratios (HRs) and 95% confidence intervals (CIs) for PFS associated with each candidate gene. Genes with $HR > 1$, such as *BUB1B*, *GAS2L3*, *NFE2L3*, *OIP5*, and *PDGFRL*, were categorized as oncogenic, as higher expression correlated with increased risk of disease progression. Conversely, genes with $HR < 1$, including *CD1D*, *CYP4B1*, *IL33*, *KLF2*, *NAPSB*, and *VILL*, were identified as tumor suppressors, with higher expression associated with improved survival outcomes. These findings underscore the dual role of the candidate genes in NPC progression and their potential as prognostic biomarkers.

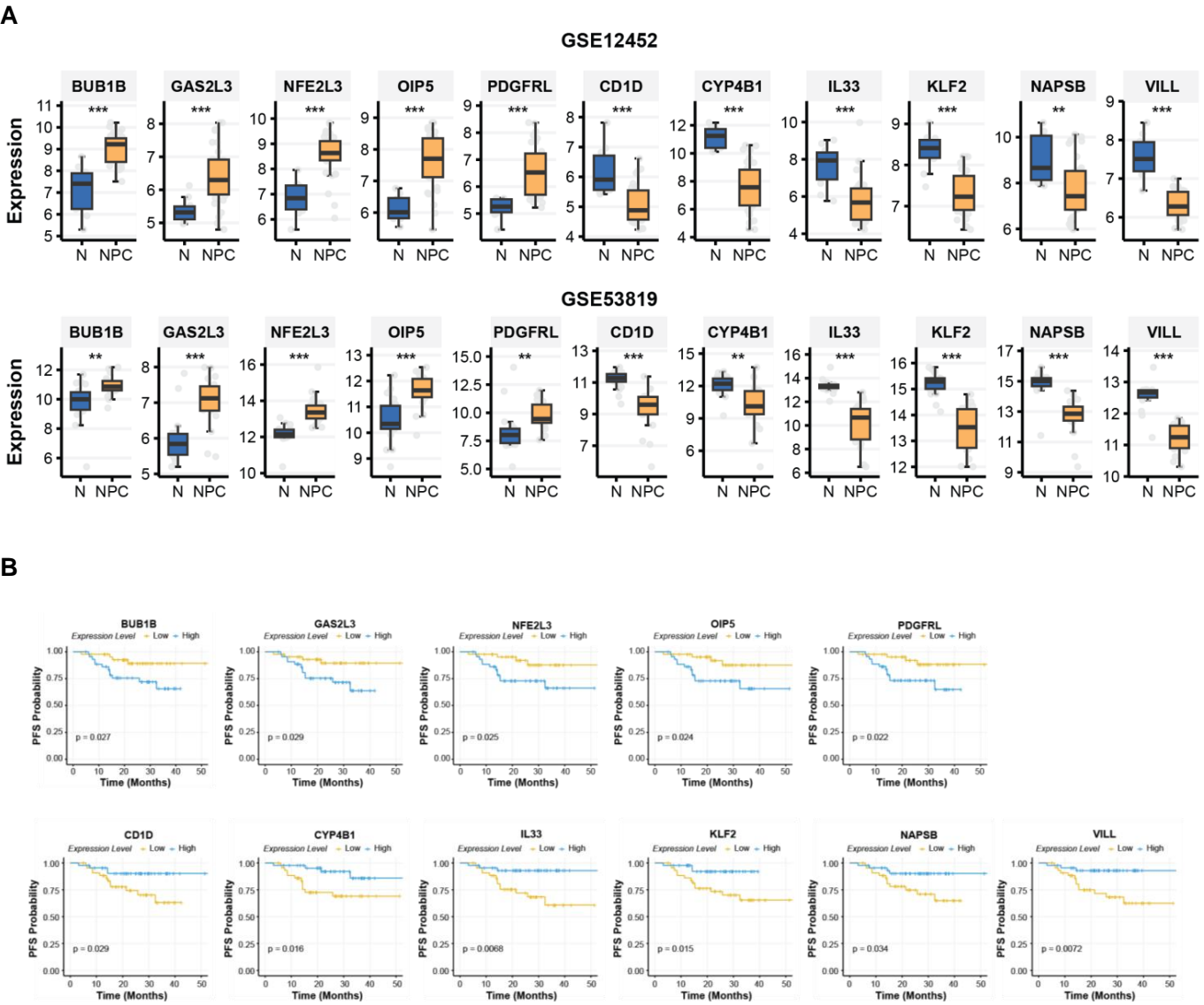


Fig. 3 contd....

C

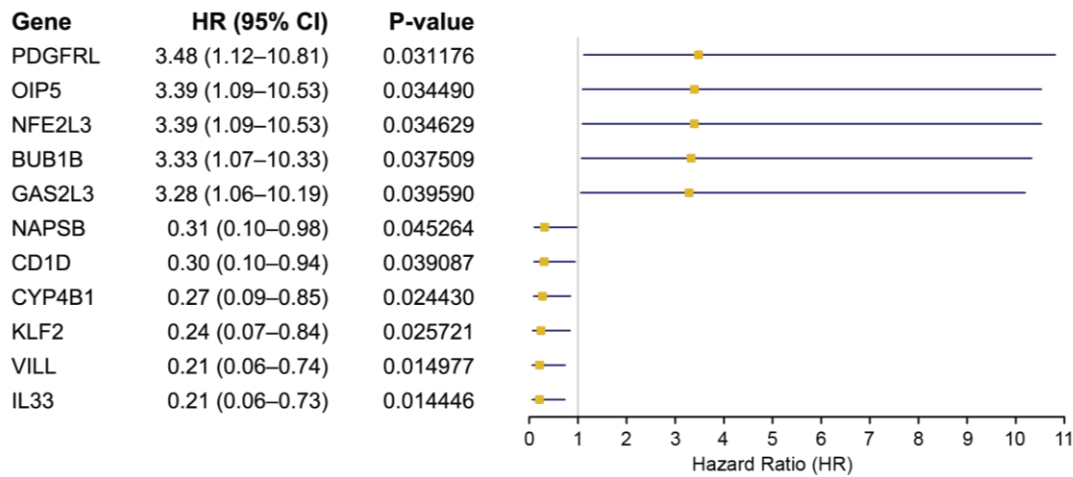


Fig. (3). Expression profiles and prognostic significance of 11 candidate genes. (A) box plot depicting differential gene expression between normal nasopharyngeal tissue (N) and nasopharyngeal carcinoma (NPC) tissue in the GSE12452 and GSE53819 datasets. Statistical significance was assessed using either the t-test or the Wilcoxon rank-sum test. $p < 0.001$ (***), $p < 0.01$ (**). (B) Kaplan-Meier survival curves for progression-free survival (PFS), stratified by high and low expression groups for each gene. Statistical significance was determined using the log-rank test. (C) Forest plot illustrating Hazard Ratios (HRs) for PFS. The forest plot displays the HRs and 95% confidence intervals (CIs) for PFS associated with each candidate gene.

3.5. RNA and Protein Correlation in Cancer Cell Lines Database

To validate the candidate biomarker genes identified from transcriptomic data (microarray and RNA-seq), we characterized their protein expression using the Cancer Cell Line Encyclopedia (CCLE) database. Among the 11 candidate genes, 10 were found in the CCLE RNA expression dataset, while one pseudogene (*NAPSB*) was not identified. Of these, 7 genes were present in the CCLE proteomic dataset, enabling RNA-protein correlation analysis. Strikingly, all 7 genes exhibited positive correlations between RNA and protein expression, indicating concordance at the transcriptomic and proteomic levels.

Six of the seven genes showed significant Pearson cor-

relations ($p < 0.05$), with correlation coefficients (r) ranging from 0.398 to 0.829. Notably, *IL33* demonstrated the strongest correlation ($r = 0.829$, $p < 0.001$), followed by *GAS2L3* ($r = 0.651$, $p < 0.001$) and *PDGFRL* ($r = 0.646$, $p < 0.001$), all of which exhibited robust RNA-protein concordance ($r > 0.5$). *VILL* ($r = 0.542$, $p < 0.001$), *BUB1B* ($r = 0.445$, $p < 0.001$), and *OIP5* ($r = 0.398$, $p < 0.001$) also showed significant correlations, albeit with slightly lower coefficients. In contrast, *NFE2L3* displayed a weak and non-significant correlation ($r = 0.142$, $p = 0.274$), suggesting potential post-transcriptional regulation or technical variability (Fig. 4). These findings highlight the strong RNA-protein concordance for most candidate genes, reinforcing their potential as robust biomarkers for further validation at both the RNA and protein levels.

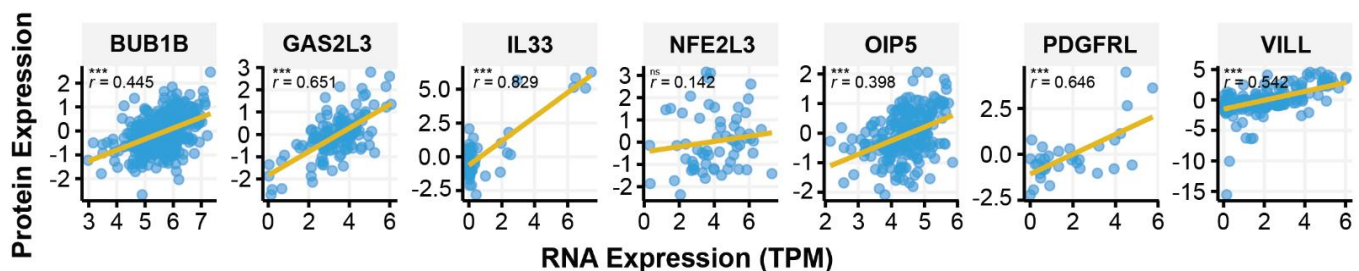


Fig. (4). RNA-Protein Correlation of Candidate Biomarker Genes in Cancer Cell Lines. Scatter plots depict the correlation between RNA expression (TPM) and protein expression levels for candidate biomarker genes in the CCLE database. Pearson correlation coefficients (r) and corresponding p-values are displayed for each gene. $p < 0.001$ (***), not-significant (ns).

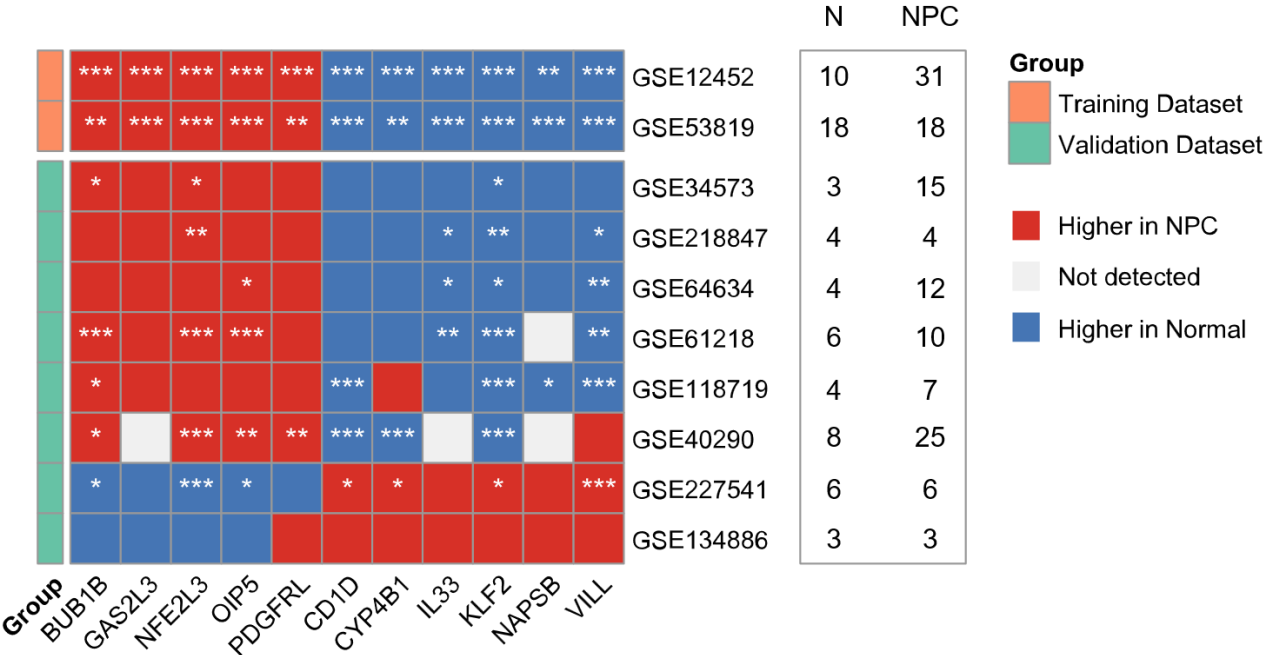


Fig. (5). Cross-dataset heatmap validation of 11 diagnostic candidate genes across eight independent GEO datasets. Red indicates higher expression in NPC, blue in normal tissues, and gray represents non-significant or unavailable data. Asterisks denote statistical significance. Sample sizes (N = normal; NPC = nasopharyngeal carcinoma) are shown alongside each dataset.

3.6. Cross-Dataset Validation Confirms Consistent Diagnostic Gene Expression Patterns

To assess the robustness of the 11 diagnostic candidate genes across independent cohorts, we performed a cross-dataset validation using eight publicly available GEO datasets that met our inclusion criteria. Among these, six datasets: GSE34573, GSE218847, GSE64634, GSE61218, GSE118719, and GSE40290, demonstrated a consistent transcriptomic pattern: *BUB1B*, *GAS2L3*, *NFE2L3*, *OIP5*, and *PDGFRL* were generally upregulated in nasopharyngeal carcinoma (NPC) samples, while *CD1D*, *CYP4B1*, *IL33*, *KLF2*, *NAPSB*, and *VILL* exhibited higher expression in normal tissues (Fig. 5). In contrast, two datasets: GSE227541 and GSE134886, showed a reversed or partially reversed pattern, which may be attributed to limited number of samples, technical variation, limited probe representation, or biological heterogeneity. Detailed statistical results are provided in Table S2.

It is important to interpret the significance levels with caution, as several datasets contained limited sample sizes per group (as indicated in Fig. 5), which reduces statistical power. Therefore, instead of relying solely on p-values, we focused on the directionality and consistency of gene expression trends across datasets, which offers a more robust validation of their diagnostic potential. Notably, due to the lack of available prognostic data in any of the external NPC datasets, independent validation of the prognostic gene signature could not be performed.

4. DISCUSSION

Our study identified 11 candidate genes from transcriptomic data, comprising 5 potential oncogenic biomarkers (*BUB1B*, *GAS2L3*, *NFE2L3*, *OIP5*, and *PDGFRL*) and 6 potential tumor suppressor biomarkers (*CD1D*, *CYP4B1*, *IL33*, *KLF2*, *NAPSB*, and *VILL*) in NPC. Notably, *NAPSB*, a pseudogene, has not been previously reported as a biomarker in NPC. However, it has been implicated in other cancers, such as hepatocellular carcinoma [24], acute myeloid leukemia [25], and pancreatic adenocarcinoma [26]. In addition, *NAPSB* were also found to be upregulated in carcinoma of the uterine cervix (CACX) [27]. Our findings suggest that *NAPSB* may play a role in NPC pathogenesis, warranting further investigation into its functional mechanisms and clinical relevance.

Among the oncogenic candidates, *BUB1B*, encoding BUB1 mitotic checkpoint serine/threonine kinase B, has been previously reported to promote NPC progression [28], a finding consistent with our results. *GAS2L3* (Growth Arrest Specific 2 Like 3), which regulates cytoskeleton organization and cytokinesis [29], has not been previously associated with NPC. Our identification of *GAS2L3* as a potential oncogenic biomarker represents a novel finding, highlighting its potential role in NPC progression. Similarly, Nuclear factor erythroid 2 (NF-E2)-related factor 3 (*NFE2L3*), a transcription factor involved in cell differentiation, oxidative stress, and tumor growth, has been implicated in various cancers such as colorectal, liver, thyroid, pancreatic, and renal cancer [30], but not

NPC. Our study is the first to report its upregulation in NPC and its potential as a prognostic biomarker. Opa interacting protein 5 (*OIP5*) was reported as a tumor promoter gene, highly expressed in NPC, and promoted NPC progression by modulating JAK2/STAT3 [31], aligning with our findings. Intriguingly, platelet-derived growth factor receptor-like (*PDGFRL*), which exhibits dual roles as a tumor suppressor in breast [32] and a tumor promoter in gastric cancer [33], has not been previously studied in NPC. Our results suggest that *PDGFRL* may serve as both a diagnostic and prognostic biomarker in NPC, underscoring its context-dependent roles in cancer biology.

Among the tumor suppressor candidates, *CD1D*, a member of the CD1 glycoprotein family, has been implicated in immune modulation within the tumor microenvironment. While *CD1D* expression is associated with aggressive renal cell carcinoma [34] It also facilitates tumor suppression through antigen presentation to NKT cells [35]. Our findings support its tumor-suppressive role in NPC, suggesting its potential as a therapeutic target. *CYP4B1*, downregulated in lung adenocarcinoma and urothelial carcinoma [36, 37] also exhibited tumor-suppressive characteristics in NPC, with low expression correlating with poor survival. Interleukin-33 (*IL33*), an alarmin cytokine involved in tissue repair, has been linked to poor progression-free survival in NPC [38, 39], consistent with our results. *KLF2*, a transcription factor downregulated in head and neck squamous cell carcinoma (HNSC) [40], has not been previously studied in NPC. Our findings confirm its tumor-suppressive role in NPC, aligning with its function in HNSC. Finally, *VILL*, which exhibits specific methylation patterns in NPC [41], emerged as a potential diagnostic and prognostic marker in our study, further validating its role in NPC pathogenesis.

Collectively, our study highlights several candidate genes with potential diagnostic and prognostic value in NPC, including *NAPSB*, which has been largely understudied in NPC. While our findings provide a transcriptomic basis for their clinical relevance, the biological functions of *NAPSB* and other identified genes in NPC pathogenesis remain to be elucidated. Further studies are warranted to explore their mechanistic roles through in vitro and in vivo functional assays, as well as validation at the protein level. Such efforts will be crucial to determine their potential utility in clinical applications, including as targets for therapy or biomarkers for early detection and prognosis.

CONCLUSION

Our study not only corroborates previous findings but also identifies novel biomarkers with potential diagnostic and prognostic significance in NPC. The dual roles of some genes, such as *PDGFRL* and *CD1D*, highlight the complexity of cancer biology and the importance of context-specific analyses. Furthermore, RNA-protein correlation analysis using the CCLE dataset revealed significant concordance between RNA and protein

expression for six genes (*BUB1B*, *GAS2L3*, *IL33*, *OIP5*, *PDGFRL*, and *VILL*). This strong RNA-protein correlation suggests that these genes could be reliably measured at either the RNA or protein level, providing flexibility in developing diagnostic assays. Future studies should focus on elucidating the functional mechanisms of these genes, particularly *GAS2L3* and *NFE2L3*, which have not been previously associated with NPC. Additionally, the pseudogene *NAPSB* warrants further investigation to determine its functional relevance in NPC. Collectively, our findings provide a robust foundation for advancing diagnostic strategies, improving prognostic stratification, and developing targeted therapies in NPC.

STUDY LIMITATIONS

This study has several limitations. First, the diagnostic analysis was based on only two datasets, which may limit the generalizability of the findings, as a larger sample size from diverse datasets could provide a more comprehensive view. For prognostic analysis, only one dataset (GSE102349) contained survival information, which may reduce the representativeness and robustness of the survival analysis. Additionally, while RNA-protein expression concordance was validated using the CCLE dataset, the observed protein expression patterns may not fully reflect the expression in human NPC tissues, due to potential differences in the tumor microenvironment and experimental conditions. The limited number of available NPC transcriptomic datasets also restricts the ability to confirm the generalizability of the findings. Finally, as an observational study based on secondary data, the results should be interpreted cautiously, and experimental validation is needed to confirm the identified biomarkers.

AUTHOR'S CONTRIBUTIONS

N.A.: Conceptualization, Methodology, Investigation, Software, Formal Analysis, Writing - Original Draft, Writing - Review & Editing;. L.R.: Conceptualization, Methodology, Software, Formal Analysis, Validation, Writing - Review & Editing;. J.Y.C.: Supervision, Conceptualization, Writing - Review & Editing;.

LIST OF ABBREVIATIONS

CCLE	=	Cancer Cell Line Encyclopedia
Cox-PH	=	Cox Proportional Hazards
DEGs	=	Differentially Expressed Genes
EBV	=	Epstein-Barr virus
GEO	=	Gene Expression Omnibus
HR	=	Hazard Ratio
NPC	=	Nasopharyngeal Carcinoma
PCA	=	Principal Component Analysis
PFS	=	Progression-free Survival
TPM	=	Transcripts Per Million

ETHICAL APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

Not Applicable.

CONSENT FOR PUBLICATION

Not applicable.

STANDARDS OF REPORTING

STROBE guidelines were followed

AVAILABILITY OF DATA AND MATERIAL

All data generated or analyzed during this study are included in this published article.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

- [1] Bray F, Laversanne M, Sung H, *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2024; 74(3): 229-63.
<http://dx.doi.org/10.3322/caac.21834>
- [2] Hsu CL, Chang YS, Li HP. Molecular diagnosis of nasopharyngeal carcinoma: Past and future. *Biomed J* 2025; 48(1): 100748.
<http://dx.doi.org/10.1016/j.bj.2024.100748> PMID: 38796105
- [3] Nicholls JM, Lee VHF, Chan SK, *et al.* Negative plasma Epstein-Barr virus DNA nasopharyngeal carcinoma in an endemic region and its influence on liquid biopsy screening programmes. *Br J Cancer* 2019; 121(8): 690-8.
<http://dx.doi.org/10.1038/s41416-019-0575-6> PMID: 31527689
- [4] Dogan S, Hedberg ML, Ferris RL, Rath TJ, Assaad AM, Chiosea SI. Human papillomavirus and Epstein-Barr virus in nasopharyngeal carcinoma in a low-incidence population. *Head Neck* 2014; 36(4): 511-6.
<http://dx.doi.org/10.1002/hed.23318> PMID: 23780921
- [5] Cho W. Nasopharyngeal carcinoma: Molecular biomarker discovery and progress. *Mol Cancer* 2007; 6(1): 1.
<http://dx.doi.org/10.1186/1476-4598-6-1> PMID: 17199893
- [6] Hong YH, Aziz N, Park JG, *et al.* The EEF1AKMT3/MAP2K7/TP53 axis suppresses tumor invasiveness and metastasis in gastric cancer. *Cancer Lett* 2022; 544: 215803.
<http://dx.doi.org/10.1016/j.canlet.2022.215803> PMID: 35753528
- [7] Tan Y, Zhou J, Liu K, *et al.* Novel prognostic biomarkers in nasopharyngeal carcinoma unveiled by mega-data bioinformatics analysis. *Front Oncol* 2024; 14: 1354940.
<http://dx.doi.org/10.3389/fonc.2024.1354940> PMID: 38854728
- [8] Sengupta S, den Boon JA, Chen IH, *et al.* Genome-wide expression profiling reveals EBV-associated inhibition of MHC class I expression in nasopharyngeal carcinoma. *Cancer Res* 2006; 66(16): 7999-8006.
<http://dx.doi.org/10.1158/0008-5472.CAN-05-4399> PMID: 16912175
- [9] Bao Y, Cao X, Luo D, *et al.* Urokinase-type plasminogen activator receptor signaling is critical in nasopharyngeal carcinoma cell growth and metastasis. *Cell Cycle* 2014; 13(12): 1958-69.
<http://dx.doi.org/10.4161/cc.28921> PMID: 24763226
- [10] Blighe K, Lun A. PCAtools: Everything principal components analysis. 2019. Available from: <https://bioconductor.org/packages/devel/bioc/vignettes/PCAtools/inst/doc/PCAtools.html>
- [11] Ritchie ME, Phipson B, Wu D, *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015; 43(7): 47.
<http://dx.doi.org/10.1093/nar/gkv007>
- [12] Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 2016; 32(18): 2847-9.
<http://dx.doi.org/10.1093/bioinformatics/btw313> PMID: 27207943
- [13] Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net. *J R Stat Soc Series B Stat Methodol* 2005; 67(2): 301-20.
<http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x>
- [14] Li L, Ching WK, Liu ZP. Robust biomarker screening from gene expression data by stable machine learning-recursive feature elimination methods. *Comput Biol Chem* 2022; 100: 107747.
<http://dx.doi.org/10.1016/j.compbiolchem.2022.107747> PMID: 35932551
- [15] Li L, Liu ZP. Detecting prognostic biomarkers of breast cancer by regularized Cox proportional hazards models. *J Transl Med* 2021; 19(1): 514.
<http://dx.doi.org/10.1186/s12967-021-03180-y> PMID: 34930307
- [16] Killian T, Gatto L. Exploiting the DepMap cancer dependency data using the depmap R package. *F1000 Res* 2021; 10: 416.
<http://dx.doi.org/10.12688/f1000research.52811.1>
- [17] Broad DepMap. DepMap 22Q4 Public. United Kingdom: Figshare 2022.10.6084/m9.figshare.21637199.v2
- [18] Fan C, Xiong F, Tang Y, *et al.* Construction of a lncRNA-mRNA Co-Expression Network for Nasopharyngeal Carcinoma. *Front Oncol* 2022; 12: 809760.
<http://dx.doi.org/10.3389/fonc.2022.809760> PMID: 35875165
- [19] Hu C, Wei W, Chen X, *et al.* A global view of the oncogenic landscape in nasopharyngeal carcinoma: An integrated analysis at the genetic and expression levels. *PLoS One* 2012; 7(7): 41055.
<http://dx.doi.org/10.1371/journal.pone.0041055> PMID: 22815911
- [20] Bo H, Gong Z, Zhang W, *et al.* Upregulated long non-coding RNA AFAP1-AS1 expression is associated with progression and poor prognosis of nasopharyngeal carcinoma. *Oncotarget* 2015; 6(24): 20404-18.
<http://dx.doi.org/10.18632/oncotarget.4057> PMID: 26246469
- [21] Liu S, Li X, Xie Q, *et al.* Identification of a lncRNA/circRNA-miRNA-mRNA network in Nasopharyngeal Carcinoma by deep sequencing and bioinformatics analysis. *J Cancer* 2024; 15(7): 1916-28.
<http://dx.doi.org/10.7150/jca.91546> PMID: 38434987
- [22] Zhang X, Song X, Lai Y, *et al.* Identification of key pseudogenes in nasopharyngeal carcinoma based on RNA-Seq analysis. *BMC Cancer* 2021; 21(1): 483.
<http://dx.doi.org/10.1186/s12885-021-08211-x> PMID: 33931030
- [23] Lin C, Zong J, Lin W, *et al.* EBV-miR-BART8-3p induces epithelial-mesenchymal transition and promotes metastasis of nasopharyngeal carcinoma cells through activating NF- κ B and Erk1/2 pathways. *J Exp Clin Cancer Res* 2018; 37(1): 283.
<http://dx.doi.org/10.1186/s13046-018-0953-6> PMID: 30477559
- [24] Ning YM, Lin K, Liu XP, *et al.* NAPSB as a predictive marker for prognosis and therapy associated with an immuno-hot tumor microenvironment in hepatocellular carcinoma. *BMC Gastroenterol* 2022; 22(1): 392.
<http://dx.doi.org/10.1186/s12876-022-02475-8> PMID: 35987606

- [25] Jian J, Wang X, Hao H, Ji C, Yuan C, Lu F. A prognostic model of pseudogenes in acute myeloid leukemia. *Clin Lab* 2023; 69(05/2023): 69. <http://dx.doi.org/10.7754/Clin.Lab.2022.220825> PMID: 37145087
- [26] Tan Z, Lei Y, Zhang B, *et al.* Analysis of immune-related signatures related to CD4+ T cell infiltration with gene co-expression network in pancreatic adenocarcinoma. *Front Oncol* 2021; 11: 674897. <http://dx.doi.org/10.3389/fonc.2021.674897> PMID: 34367961
- [27] Roychowdhury A, Samadder S, Das P, *et al.* Deregulation of H19 is associated with cervical carcinoma. *Genomics* 2020; 112(1): 961-70. <http://dx.doi.org/10.1016/j.ygeno.2019.06.012> PMID: 31229557
- [28] Qin LT, Huang SW, Huang ZG, *et al.* Clinical value and potential mechanisms of BUB1B up-regulation in nasopharyngeal carcinoma. *BMC Med Genomics* 2022; 15(1): 272. <http://dx.doi.org/10.1186/s12920-022-01412-8> PMID: 36577966
- [29] Sharaby Y, Lahmi R, Amar O, *et al.* Gas2l3 is essential for brain morphogenesis and development. *Dev Biol* 2014; 394(2): 305-13. <http://dx.doi.org/10.1016/j.ydbio.2014.08.006> PMID: 25131197
- [30] Xiong G, Li J, Yao F, Yang F, Xiang Y. New insight into the CNC-bZIP member, NFE2L3, in human diseases. *Front Cell Dev Biol* 2024; 12: 1430486. <http://dx.doi.org/10.3389/fcell.2024.1430486> PMID: 39149514
- [31] Zheng YQ, Cui YR, Yang S, Wang YP, Qiu YJ, Hu WL. Opa interacting protein 5 promotes metastasis of nasopharyngeal carcinoma cells by promoting EMT via modulation of JAK2/STAT3 signal. *Eur Rev Med Pharmacol Sci* 2019; 23(2): 613-21. PMID: 30720169
- [32] Kawata K, Kubota S, Eguchi T, *et al.* A tumor suppressor gene product, platelet-derived growth factor receptor-like protein controls chondrocyte proliferation and differentiation. *J Cell Biochem* 2017; 118(11): 4033-44. <http://dx.doi.org/10.1002/jcb.26059> PMID: 28407304
- [33] Yang Q, Li X, Zhu W. Identification of a unique stress response state of T cells-related gene signature in patients with gastric cancer. *Aging* 2024; 16(11): 9709-26. <http://dx.doi.org/10.18632/aging.205895> PMID: 38848147
- [34] Chong TW, Goh FY, Sim MY, *et al.* CD1d expression in renal cell carcinoma is associated with higher relapse rates, poorer cancer-specific and overall survival. *J Clin Pathol* 2015; 68(3): 200-5. <http://dx.doi.org/10.1136/jclinpath-2014-202735> PMID: 25477528
- [35] Li Y, Zhao C, Liu J, *et al.* CD1d highly expressed on DCs reduces lung tumor burden by enhancing antitumor immunity. *Oncol Rep* 2019; 41(5): 2679-88. <http://dx.doi.org/10.3892/or.2017.5544> PMID: 30864713
- [36] Liu X, Jia Y, Shi C, *et al.* CYP4B1 is a prognostic biomarker and potential therapeutic target in lung adenocarcinoma. *PLoS One* 2021; 16(2): 0247020. <http://dx.doi.org/10.1371/journal.pone.0247020> PMID: 33592039
- [37] Lin JT, Chan TC, Li CF, *et al.* Downregulation of the cytochrome P450 4B1 protein confers a poor prognostic factor in patients with urothelial carcinomas of upper urinary tracts and urinary bladder. *Acta Pathol Microbiol Scand Suppl* 2019; 127(4): 170-80. <http://dx.doi.org/10.1111/apm.12939> PMID: 30803053
- [38] Dai Y, Chen W, Huang J, *et al.* Identification of key pathways and genes in nasopharyngeal carcinoma based on WGCNA. *Auris Nasus Larynx* 2023; 50(1): 126-33. <http://dx.doi.org/10.1016/j.anl.2022.05.013> PMID: 35659152
- [39] Arslan İ, Yılmazçoban H, Eyigör H, *et al.* The effect of interleukin-33 expression on prognosis in patients with nasopharyngeal carcinoma. *Acta Otorrinol Esp* 2025; 76(2): 76-82. <http://dx.doi.org/10.1016/j.otoeng.2024.01.011>
- [40] Xu R, Chen Y, Wei S, Chen J. Comprehensive pan-cancer analysis of the prognostic role of KLF transcription factor 2 (KLF2) in human tumors. *OncoTargets Ther* 2024; 17: 887-904. <http://dx.doi.org/10.2147/OTT.S476179> PMID: 39507409
- [41] Fu XY, Zhou ZY, Yang TY, *et al.* Plasma cfDNA VILL gene methylation as a diagnostic marker for nasopharyngeal carcinoma. *Clin Epigenetics* 2025; 17(1): 38. <http://dx.doi.org/10.1186/s13148-025-01847-7> PMID: 40016817

DISCLAIMER: The above article has been published, as is, ahead-of-print, to provide early visibility but is not the final version. Major publication processes like copyediting, proofing, typesetting and further review are still to be done and may lead to changes in the final published version, if it is eventually published. All legal disclaimers that apply to the final published article also apply to this ahead-of-print version.