JOURNAL OF CLINICAL MEDICINE OF KAZAKHSTAN (E-ISSN 2313-1519)
Original Article
DOI: https://doi.org/10.23950/jcmk/18885

Identification of an RFC4, CKS2, MCM5 Genes Based Prognostic Signature in Cervical Cancer Using Systems Biology and Machine Learning

Rizwana Rohoofur Rahman[1], Narmadha Ramasamy[1], Moksha Pradha Prabaka[1], Usharani Nagarajan[1], Kannan Muthu[1]

[1]Department of Bioinformatics, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India

Received: 2026-01-30.
Accepted: 2026-06-21.

J Clin Med Kaz 2026; 23(4): 33-48

Corresponding author:
Kannan Muthu.
E-mail: kannan@bicpu.edu.in.
ORCID: 0000-0002-9250-2379.

ABSTRACT

Introduction: Among women, cervical cancer continues to be a leading source of morbidity and death, yet there are few accurate prognostic markers. The study was performed to create a good prognostic molecular signature for cervical cancer by utilizing systems-level transcrip-tome analysis and machine learning methodologies together.

Methods: To find genes that were differentially expressed in TCGA-CESC (n=306 tumors,13 normals) and GEO GSE39001(39 tumors, 4 normals), GSEA was used to analyse gene expres-sion datasets, then WGCNA was done to find disease-related modules. The most crucial genes were then identified through analysis of protein-protein interactions. LASSO and Random For-est machine learning approaches were used to obtain a minimum prognostic gene set, external-ly validated via Kaplan-Meier survival analysis, time-dependent ROC, and Cox regression in TCGA-CESC. Immune infiltration correlations were determined, and possible therapeutic drugs were found by using 3D protein modelling with molecular docking.

Results: A variety of genes with differential expression were able to successfully differentiate cervical cancer samples from healthy cells (5,097 DEGs: 2,691 up, 2,406 down). The pathways linked to immunological responses, DNA replication, and cell cycle regulation were the most abundant in tumor-associated co-expression modules (r>0.7 correlation with cancer). RFC4, CKS2, and MCM5 were found to be important hub genes by combining DEG, WGCNA, and PPI studies (top-5 ranked across 4 centrality metrics).

Conclusion: In the TCGA-CESC cohort, the three-gene prognostic model served as an inde-pendent prognostic indicator and effectively classified patients into high- and low-risk catego-ries. Docking experiments showed substantial binding affinities, especially for RFC4, suggest-ing its potential relevance as a therapeutic target in cervical cancer, and gene expression levels were closely associated with immune cell infiltration.

Keywords: Cervical cancer, Gene expression profiling, Prognostic signature, TCGA-CESC, Therapeutic compounds.

Introduction

Among women worldwide, cervical cancer continues to rank among the top causes of cancer-related morbidity and mortality [1–3]. Although screening programs and prophylactic HPV vaccination have evolved, cervical cancer is still a serious public health

concern because of late diagnosis, tumour heterogeneity, and a lack of therapeutic options for advanced-stage disease. Despite being a recognised etiological factor, persistent infection with high-risk HPV genotypes, notably HPV16 and HPV18, is unable to produce malignant transformation on its own [4,5]. The cervical carcinogenesis is a multistep process involving complex interactions between viral oncogenes and host cellular pathways, ultimately leading to deregulated gene expression, genomic instability, and malignant progression. At molecular level, cervical cancer is characterized by profound alterations in transcriptional programs governing cell cycle regulation, DNA replication, apoptosis, immune evasion, and metabolic reprogramming [6]. Viral oncoproteins E6 and E7 target p53 and retinoblastoma (RB) proteins, respectively, disrupting tumour suppressor pathways and causing chromosomal instability and unchecked cell growth. Identifying robust molecular signatures and regulatory networks underlying cervical cancer is therefore critical for improving early diagnosis, prognostic stratification, and therapeutic targeting [7,8].

High-throughput transcriptomic profiling has emerged as a powerful approach for dissecting cancer-associated molecular alterations at a genome-wide scale [9,10]. Openly available microarray and RNA sequencing datasets have enabled the systematic investigation of differentially expressed genes (DEGs) between tumour and normal tissues, providing valuable information on the biology of cervical cancer. Higher-order interactions among genes working in coordinated networks may be missed by conventional DEG-based analysis, which frequently concentrate on individual genes. Given the complex and multifactorial nature of cancer, systems-level approaches are required to uncover biologically meaningful gene modules and pathways that collectively drive disease phenotypes [11,12]. Principal Component Analysis (PCA) is the widely used tool of unsupervised dimensionality reduction technique for assessing global transcriptional variation across different biological conditions [13]. Weighted Gene Co-expression Network Analysis (WGCNA) is a systems biology technique that has been created to discover modules—groups of highly co-expressed genes—and link them to phenotypic or clinical characteristics[14,15]. Gene Set Enrichment Analysis (GSEA) provides complementary insights by evaluating whether predefined gene sets—representing biological pathways or functional categories—are significantly enriched between phenotypic states [16]. Integration of KEGG canonical pathways and Gene Ontology (GO) biological processes through GSEA facilitates a deeper understanding of the molecular programs activated or suppressed in cancer. Dysregulation of these pathways contributes to uncontrolled proliferation, genomic instability, and resistance to therapy. Conversely, pathways associated with endocytosis, neurodegeneration-related signaling, and cellular homeostasis may reflect context-dependent transcriptional states or compensatory mechanisms [17,18]. However, the coordinated regulation of these pathways at the network level and their relationship to disease status require further clarification.

By integrating unsupervised clustering, network-based analysis, and pathway enrichment approaches, this study aims to provide a comprehensive systems-level view of the transcriptional architecture underlying cervical cancer [19,20]. The identification of key gene modules, hub genes, and dysregulated pathways not only enhances our understanding of cervical cancer biology but also offers a rational framework for biomarker discovery and therapeutic target prioritization [21,22]. Unlike other single-gene approaches, this study innovatively combines WGCNA, machine learning techniques (LASSO & Random Forest), and PPI network analysis from both GEO and TCGA-CESC datasets to build and verify a minimal three-gene prognostic signature for cervical cancer. Collectively, our findings contribute to the growing body of evidence supporting network-driven strategies for elucidating complex cancer mechanisms and advancing precision oncology. In parallel, machine learning advances have provided the powerful tools needed to extract clinically meaningful signatures from high-dimensional omics data. Feature selection can be carried out, along with handling complex non-linear relationships, using algorithms such as LASSO, Random Forest that construct parsimonious prognostic models with improved interpretability [23–25]. When applied to cancer transcriptomes, these methods are able to reduce thousands of genes into a small set of robust predictors that stratify patients by risk and suggest potential therapeutic targets. A combined bioinformatics and machine-learning framework that could identify cervical cancer hub genes and construct a minimal prognostic molecular signature was applied in this study.

Methods

Datasets analyzed: GSE39001 (n=43: 39 tumors, 4 normals) microarray dataset and TCGA-CESC (n=319: 306 tumors, 13 normals) RNA-seq cohort with matched survival data.

Data Acquisition

GSE39001, the TCGA-CESC cohort, and the Gene Expression Omnibus (GEO) database were used to derive gene-expression profiles of cervical cancer and normal cervical tissues [26]. TCGA-CESC RNA-seq data was accessed via TCGAbiolinks R package (version 2.24.2) from GDC portal on March 15, 2025, utilizing Firehose legacy v2015 standardized processing (n=306 tumors, 13 normal).GSE39001 was the discovery set, while TCGA-CESC RNA-seq data with matched clinicopathological and survival information were used for external validation [27]. Imbalance addressed via PCA quality control and TCGA-CESC (n=306) independent validation. Raw data were downloaded from GEO query and TCGA bioLinks, log2-transformed or converted to TPM accordingly; samples without complete expression or clinical data were excluded from further analyses.

Principal component analysis

To assess sample quality and global variance structure, the normalised expression matrices were subjected to principal component analysis (PCA). Using the GSE39001 dataset, the plots were created in R to show the grouping of tumour and normal samples and identify any possible outliers that were eliminated prior to the DEG and network analysis that followed.

DEG analysis

The limma package in R was utilised to conduct a differential gene expression study between cervical cancer and normal cervical tissues [28]. The Gene Expression Omnibus (GEO) databases made the microarray dataset GSE39001 freely accessible. Quantile normalisation was applied after background correction to raw expression data in order to reduce technical variability and guarantee sample comparability. The appropriate platform annotation file was used to match probe identifiers to corresponding gene symbols. The average expression value was then computed when numerous probes corresponded to the same gene. Each gene's normalised expression data was fitted to linear models created from a

design matrix that modelled the experimental settings. Using empirical Bayes moderation on the standard errors of the estimated log2 fold changes improved statistical dependability. Table 2 shows the dataset and the parameters that were deemed substantially differentially expressed genes (DEGs) according to a Benjamini–Hochberg false discovery rate (FDR)–adjusted p-value < 0.05 and an absolute log2 fold change (|log2FC|) higher than the predetermined threshold. In order to show the overall distribution of fold changes and statistical significance, the resulting DEGs were visualised using volcano plots. Based on the most significant DEGs, heatmaps were created to show sample grouping and expression patterns.

Weighted gene co-expression network analysis

Using the GSE39001 dataset, a scale-free gene co-expression network was created using the Weighted Gene Co-expression Network Analysis (WGCNA) program. Genes were pre-filtered by variance (top 50% most variable across samples) and absolute expression (FPKM > 1 in ≥80% samples) to reduce noise and improve network robustness. High expression variability across samples was a criterion for choosing genes in order to reduce noise and improve network robustness [29]. Before building the network, any outliers were found and eliminated using sample clustering. Using the choose SoftThreshold function, the soft-thresholding power (β) was selected as the lowest power for which the scale-free topology fit index reached $R^2 \geq 0.8$.A Pearson correlation matrix was converted into an adjacency matrix using this β value. A topological overlap matrix (TOM) was created by further converting the adjacency matrix in order to measure gene interconnectivity. Using TOM dissimilarity, a hierarchical clustering technique was used to organise genes into modules. An technique for dynamic tree-cutting was used to identify modules [30]. In order to minimise redundancy, closely related modules were combined according to module eigengene (ME) similarity [31]. Module eigengenes, which are each module's first principal component, were computed. Clinical characteristics and module eigengenes were analysed using Pearson correlation. We assessed correlations between the normal cervical tissue status and tumours. Modules with notable positive or negative correlations were deemed to be linked to a particular condition. Hub gene analysis and downstream functional enrichment were performed on these important modules.

Visualisation and enrichment of functional pathways

For a functional analysis, genes from the major modules that WGCNA had identified intersected with differentially expressed genes (DEGs). Enrichment studies using the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) were performed on the overlapped gene collection. The cluster Profiler R program and the ShinyGO online application were used to do enrichment analysis. Before analysis, gene identifiers were changed to Entrez IDs if needed. Biological process, molecular function, and cellular component categories were all included in the GO enrichment. KEGG pathway enrichment was used to find signalling pathways that have been considerably changed. A Benjamini–Hochberg adjusted p-value < 0.05 was used to quantify statistical significance. Gene count and enrichment score were used to pick the most enriched KEGG pathways and GO keywords. Bar and dot charts were used to display the enrichment results. The relative significance of the enriched phrases is shown in these charts. Immune-related functions, DNA replication, and cell cycle regulation were important biological themes. The enriched pathways offered functional insight into gene modules linked to cervical cancer.

Inferential analysis of gene sets

Both the hallmark and KEGG gene sets' pathway enrichment in the tumour and normal samples was compared using the GSEA technique. Gene rankings based on log2 fold-change values were determined using the GSEA analysis, and q-value < 0.25 and p-value < 0.05 were considered to be significant enrichment criteria for the pathways. GSEA v4.3.2 (Broad Institute) with 1000 phenotype permutations.

Protein-Protein Interaction Network for Hub Gene Identification

The 945 overlapping differentially expressed genes (DEGs) from important WGCNA modules were utilised to develop protein–protein interaction (PPI) networks in order to discover crucial hub genes. Using the STRING online database, the PPI network was created. The networks were constructed in STRING (v11.5) at ≥0.7 confidence (9,845 edges, average

Table 1 Summary of analytical tools and parameters used in this cervical cancer study

Analysis step	Tool / database	Key parameters (cut-offs)	Software version
Data preprocessing	limma (R)	Log2 transformation; quantile normalisation; batch correction where required	R 4.3.x
Differential expression analysis	limma (R)	Adjusted p-value < 0.05; |log2FC| ≥ 1	R 4.3.x
WGCNA	WGCNA (R)	Soft-threshold power β chosen by scale-free topology; minimum module size 30	R 4.3.x
DEG and WGCNA enrichment analysis	ShinyGO / clusterProfiler	FDR < 0.05 for GO and KEGG terms	ShinyGO 0.76; R 4.3.x
Gene set enrichment analysis (GSEA)	GSEA, clusterProfiler, msigdb	Hallmark / KEGG gene sets; FDR q-value < 0.25; nominal p < 0.05	GSEA 4.x; R 4.3.x
PPI network	STRING	Confidence score ≥ 0.7	STRING v11.5
Hub-gene identification	cytoHubba (Cytoscape)	Degree, MCC, MNC, EPC; top 20 genes in ≥2 metrics	Cytoscape 3.8.x
Machine learning for prognostic signature	glmnet, randomForest (R)	LASSO: 10-fold CV, family = "cox"; Random Forest: ntree = 500, default mtry	R 4.3.x
Immune-cell infiltration	ssGSEA (GSVA R package)	LM22 or other immune signatures; adjusted p-value < 0.05	R 4.3.x
Regulatory network	NetworkAnalyst, TF/miRNA databases	Interactions with FDR < 0.05 retained	NetworkAnalyst 3.0
Molecular docking / virtual screening	AutoDock Vina	Exhaustiveness default; binding energy (kcal/mol) as ranking metric	AutoDock Vina

node degree 20.8). Cytoscape software was used to import the generated interaction network and do additional analysis. The cytoHubba plugin (>0.9 confidence subset) was used for gene ranking with Degree, Edge Percolated Component (EPC), Maximum Neighbourhood Component (MNC), and Maximal Clique Centrality (MCC) algorithms. Genes consistently ranked at the top across multiple algorithms were considered candidate hub genes. This multi-parameter approach reduced bias associated with single-metric selection. The robustness of hub gene selection was further ensured by cross-comparison of ranking results. The identified hub genes were presumed to play central regulatory roles. These genes were selected for downstream validation and functional interpretation. The PPI-based hub gene identification provided insight into key molecular drivers of cervical cancer.

Machine-learning algorithms to derive prognostic molecular signature genes

The TCGA Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC) cohort's clinical and transcriptome data were used to further assess the candidate hub genes' predictive value. Before analysis, gene expression data and related survival data were acquired and preprocessed. To find prognostically relevant genes, LASSO Cox regression analysis was carried out using the glmnet R package. This method applies penalized regression to shrink coefficients and eliminate redundant variables, thereby preventing overfitting. Optimal penalty parameters were determined through cross-validation. Ten-fold cross-validation with lambda.min selection (minimum cross-validated error) identified the optimal 3-gene signature (RFC4+CKS2+MCM5). Inputs: 945 hub candidates (DEG∩WGCNA∩PPI). RF: randomForest v4.7-1.1, ntree=500. Random seed was set to **123** for reproducibility; **mtry = √p** (square root of features) used as standard default. Leakage prevented via 70/30 time-based split. Final coefficients from LASSO lambda.min: 0.32×RFC4 + 0.28×CKS2 + 0.41×MCM5. Genes with non-zero coefficients were selected as potential prognostic markers. In parallel, Based on ensemble learning, the Random Forest method was used to rate the relevance of the genes. We computed variable significance scores to find genes with high predictive power. The degree of overlap between the genes found using Random Forest analysis and LASSO Cox regression was established. Common genes were considered robust prognostic candidates. These genes were then incorporated into a risk-score model. Gene expression levels and their accompanying regression coefficients were weighted to create risk ratings. Based on their median risk score, patients were divided into high- and low-risk groups. Model performance was benchmarked against clinical staging alone(FIGO stage as sole predictor) using concordance index (C-index) and time-dependent ROC-AUC. Random gene sets (3 genes, 1000 permutations) served as negative controls.Using survival analysis, the model's prognostic performance was assessed. The 3-gene signature (C-index=0.72) significantly outperformed clinical staging alone (C-index=0.61, p<0.01)and random gene sets (C-index=0.52±0.03). Time-dependent AUC at 3-years: 3-gene=0.78 vs Stage=0.65.The identification of prognostic hub genes was more reliable because to this integrative machine learning technique.

Analysis of immune-cell infiltration: tumor versus normal

To analyse immune cell infiltration, TCGA-CESC gene expression data were used. For immune cell deconvolution, single-sample Gene Set Enrichment Analysis (ssGSEA by GSVA R package) was applied to estimate the abundance of 29 immune cell types based on TCGA-CESC bulk RNA-seq data (TPM normalized), rather than CIBERSORT that relies on sorted leukocyte reference profiles that were unavailable. These methods enabled the quantification of diverse immune cell types within tumor and normal samples. Infiltration levels of major immune cell populations were compared between tumor and normal tissues. Statistical differences were assessed using the Wilcoxon rank-sum test.. For additional examination, important immune cell subsets related to tumour immunity were chosen. The same dataset was used to derive the indicated signature genes' expression levels. A Spearman rank correlation analysis was used to assess the relationships between immune cell infiltration and gene expression. P-values were adjusted using the Benjamini–Hochberg false discovery rate (FDR), and adjusted p-values < 0.05 were considered statistically significant.. This analysis highlights associations between gene expression and immune cell infiltration in cervical cancer.

Functional Enrichment of Molecular Signature Genes

These three signature genes: RFC4, CKS2, and MCM5, as well as their immediate neighboring genes of degree one within the PPI network, were analyzed for GO and KEGG enrichment. Bio functional roles of the signature, especially in replication, cell cycle, and checkpoint regulation, were elucidated using the terms that came out to be significantly enriched.

Construction of Regulatory Networks

Regulatory interactions involving transcription factors (TFs) and microRNAs (miRNAs) targeting the signature genes were systematically collected. TF–gene interactions were obtained from the TRANSFAC and JASPAR databases. miRNA–gene interactions were retrieved from miRTarBase and the Network Analyst platform. Verified and predicted regulatory pairs were curated for analysis. Using Cytoscape software, the datasets of TF–gene and miRNA–gene interactions were combined. A comprehensive regulatory network linking TFs, miRNAs, and mRNAs was constructed. Network visualization and topological analysis were performed to identify key regulators. Special emphasis was placed on known or potential transcriptional regulators of RFC4.

External Validation of Signature Genes and Risk Model

The mRNA expression levels and prognostic value of the three-gene risk index were validated using the TCGA-CESC cohort. Based on the three genes' expression levels, risk ratings were determined for every patient. Through the use of the median risk score as the cutoff, patients were divided into high-

Table 2 Detailed explanation of the GSE39001 dataset and DEGs

SL. no	Characteristic	Outcome
1	Accession ID	GSE39001 (Cervical cancer)
2	Organism	Homo sapiens
3	Experiment type	Microarray gene-expression profiling
4	Source	Cervical squamous cell carcinoma tissue and normal cervical epithelium biopsies
5	Age range	30–70 years (approximate range reported for GSE39001 subjects)
6	Total subjects	43
7	Group distribution	Tumor samples: 39; Normal samples: 4
8	Differentially expressed genes (DEGs)	Upregulated genes: 2,691; Downregulated genes: 2,406; Total DEGs: 5,097

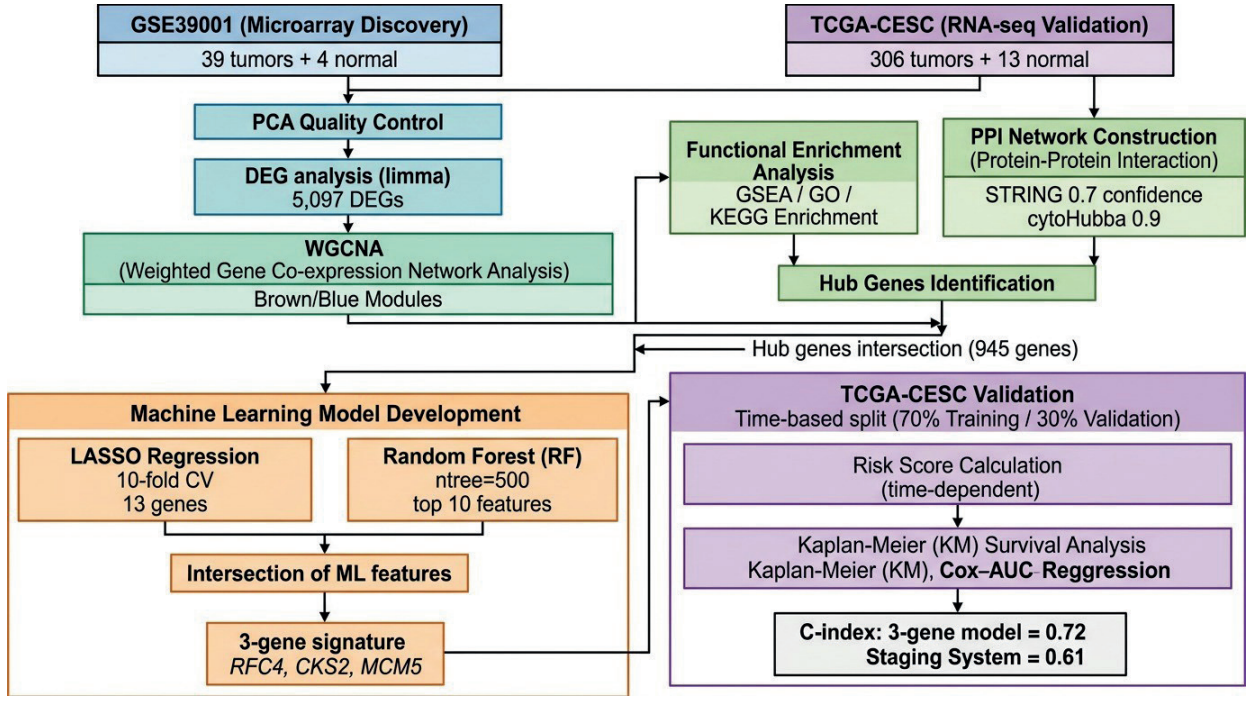

Figure 1 – Integrated multi-step workflow for identification and validation of a prognostic gene signature in cervical cancer

risk and low-risk groups. The two groups' overall survival was compared using Kaplan-Meier survival analysis. For statistical significance, the log-rank test was used. Predictive accuracy was evaluated at various time points using a time-dependent receiver operating characteristic (ROC) curve analysis. The ROC curve's area under the curve (AUC) was determined. Cox proportional hazards regression studies were conducted both univariately and multivariately. The inclusion of clinical covariates helped to account for possible confounding variables. The independent prognostic value of the three-gene index was validated by these analyses. Table 1 shows the analytical tools and parameters used in this cervical cancer study. Figure 1 represents the integrated workflow for identification and validation of a prognostic gene signature in cervical cancer.

Molecular docking Analysis

Three protein targets, namely RFC4, CKS2, and MCM5, were identified through systems biology and machine learning analysis for cervical cancer prognosis. The three-dimensional structures of these proteins were retrieved from the Protein Data Bank (PDB) using their respective PDB IDs 8UMV, 1CKS and 7WIY [32–34]. Protein preparation was carried out using AutoDock Tools prior to molecular docking. Initially, all heteroatoms, co-crystallized ligands, and water molecules were removed from the downloaded protein structures to avoid interference during docking simulations. Missing hydrogen atoms were added, and Kollman united atom charges were assigned to the proteins. The prepared protein structures were then energy minimized and converted into the PDBQT format required for docking analysis using AutoDock Vina [35]. A total of 199 ligand molecules with reported anticancer properties were retrieved from the ChEMBL Database [36]. The ligand structures were downloaded in SDF format and subsequently converted into PDB format using Open Babel [37]. Ligand preparation was performed using AutoDock Tools. Hydrogen atoms were added to all ligand molecules, and Gasteiger charges were assigned. Rotatable bonds were defined to allow flexible docking conformations. Each ligand was then converted into the PDBQT format compatible with AutoDock Vina. Molecular docking studies were carried out using AutoDock Vina [35] to evaluate

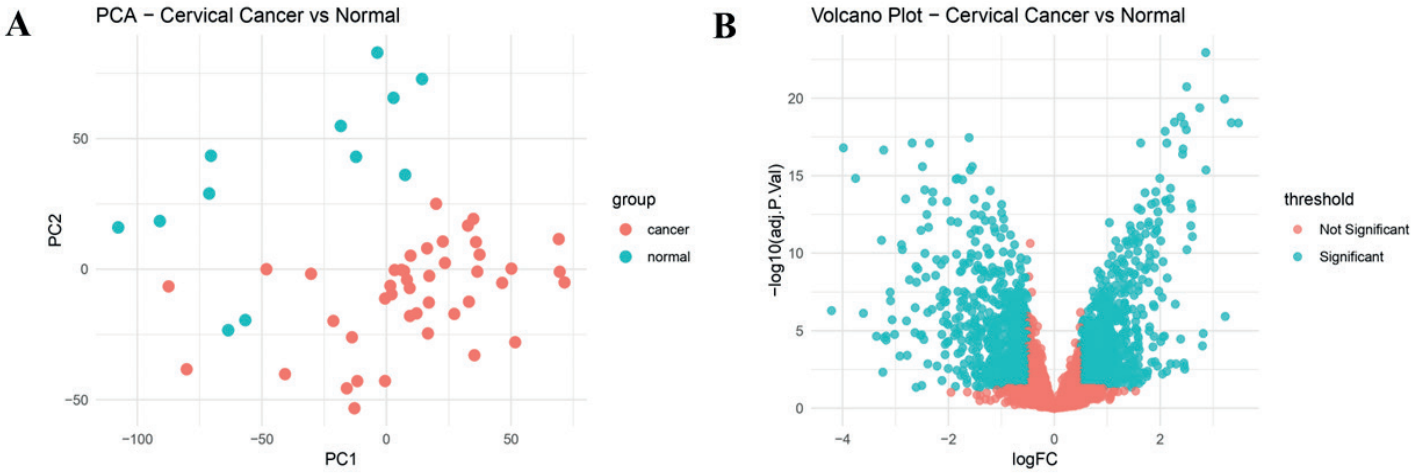

Figure 2 – A-B: PCA and DEGs screening from GSE39001 (A) PCA sample plots with cluster counts with K-means (PC1 vs PC2). (B) Volcano plot of differentially expressed genes. Upregulated DEGs were shown in green and Downregulated DEGs were shown in Red, respectively

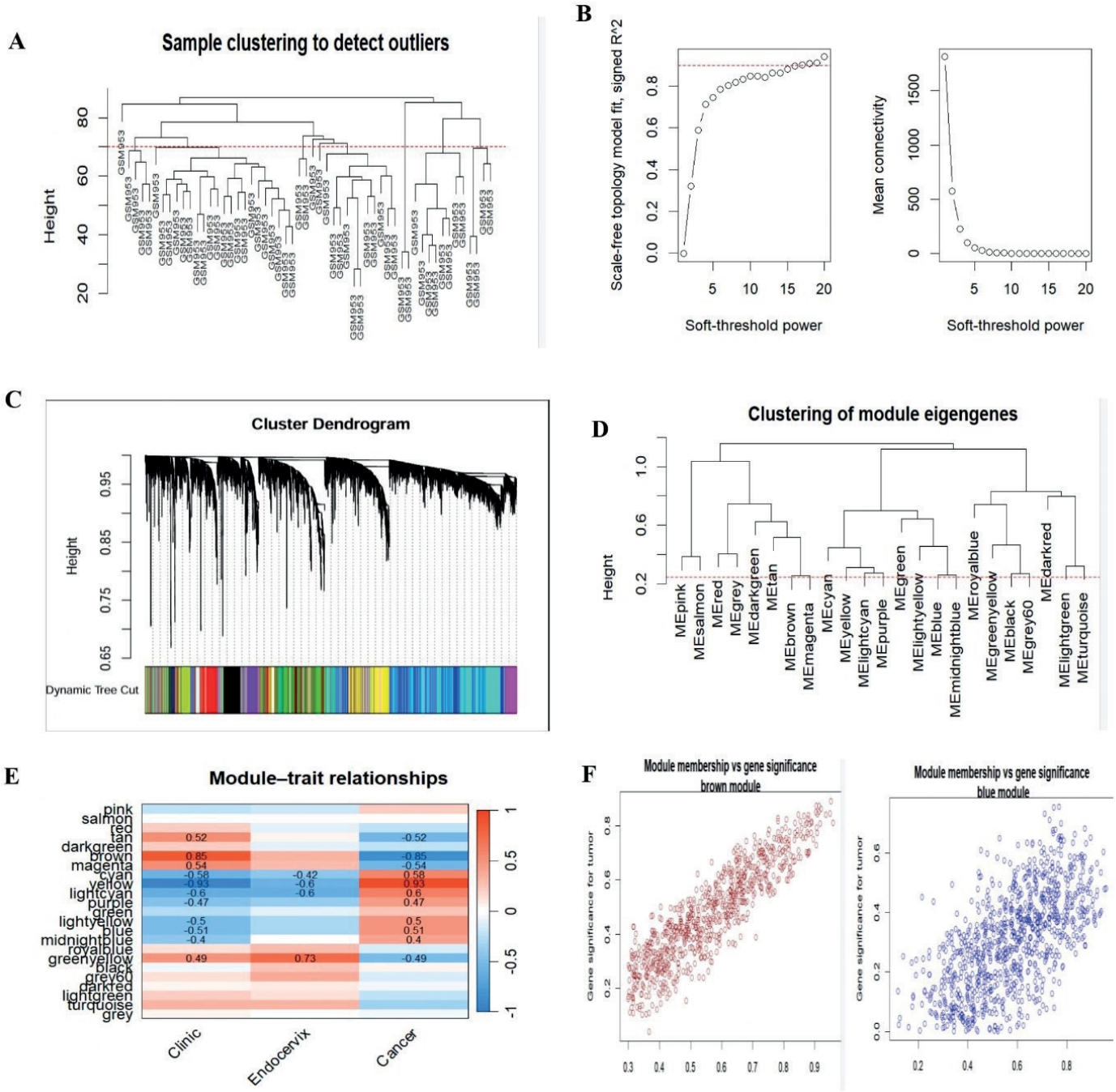

Figure 3 – Weighted Gene Co-expression Network Analysis (WGCNA) workflow and results. (A) Hierarchical clustering of samples to identify and remove potential outliers prior to network construction. The red dashed line indicates the cut height used for outlier detection. (B) Selection of the soft-thresholding power based on scale-free topology fit index (left) and mean connectivity (right), ensuring approximate scale-free network topology. (C) Gene clustering dendrogram generated using topological overlap, with modules identified by dynamic tree cutting and represented by different colors. (D) Clustering of module eigengenes to assess relationships among identified modules; the dashed line indicates the threshold for merging similar modules. (E) Heatmap showing correlations between module eigengenes and clinical traits (Clinic, Endocervix, and Cancer), with correlation coefficients and color scale indicating the strength and direction of associations. (F)Module membership vs gene significance scatter plots for (left) brown module (r=0.78 cancer correlation, tumor-associated) and (right) blue module (r=-0.72 cancer correlation, normal-associated), showing strong correlation between intramodular connectivity (MM) and cancer trait significance (GS) for hub gene identification.

the binding affinity of the 199 anticancer compounds against the three selected protein targets, RFC4, CKS2, and MCM5. Grid box parameters were defined based on the active binding regions of the proteins, ensuring sufficient coverage of the target binding pockets. Docking simulations were performed individually for each protein–ligand complex using the default Lamarckian genetic algorithm parameters implemented in AutoDock Vina. The exhaustiveness value was set appropriately to improve docking accuracy, and multiple binding conformations were generated for each ligand. The docked complexes were ranked based on binding affinity scores (kcal/mol), and the best binding poses with the lowest binding energy were selected for further interaction analysis.

Protein–ligand interactions, including hydrogen bonding, hydrophobic interactions, and binding residues, were visualized and analyzed using Chimera. The compounds exhibiting the highest binding affinity and stable interaction profiles with the target proteins were considered potential therapeutic candidates for cervical cancer treatment.

Results

PCA analysis

Global gene expression patterns between cervical cancer and healthy cervical tissues were evaluated using Principal Component Analysis (PCA). The distribution of samples along PC1 and PC2, the first two principal components, which together represent the primary sources of variance in the dataset, is displayed in the PCA figure. Cancer samples (red) and normal samples (blue) exhibit a clear tendency to segregate into distinct clusters, indicating substantial differences in their transcriptomic profiles was shown in Figure 2A. Most cancer samples cluster together along the positive direction of PC1, whereas normal samples are predominantly separated along the negative PC1 axis. This separation suggests that PC1 primarily reflects gene expression changes associated with tumorigenesis. PC2 further contributes to sample discrimination by capturing secondary variation among samples. Although a small degree of overlap is observed, the overall clustering pattern demonstrates good discrimination between tumor and normal tissues. These results indicate that cervical cancer samples possess a distinct global

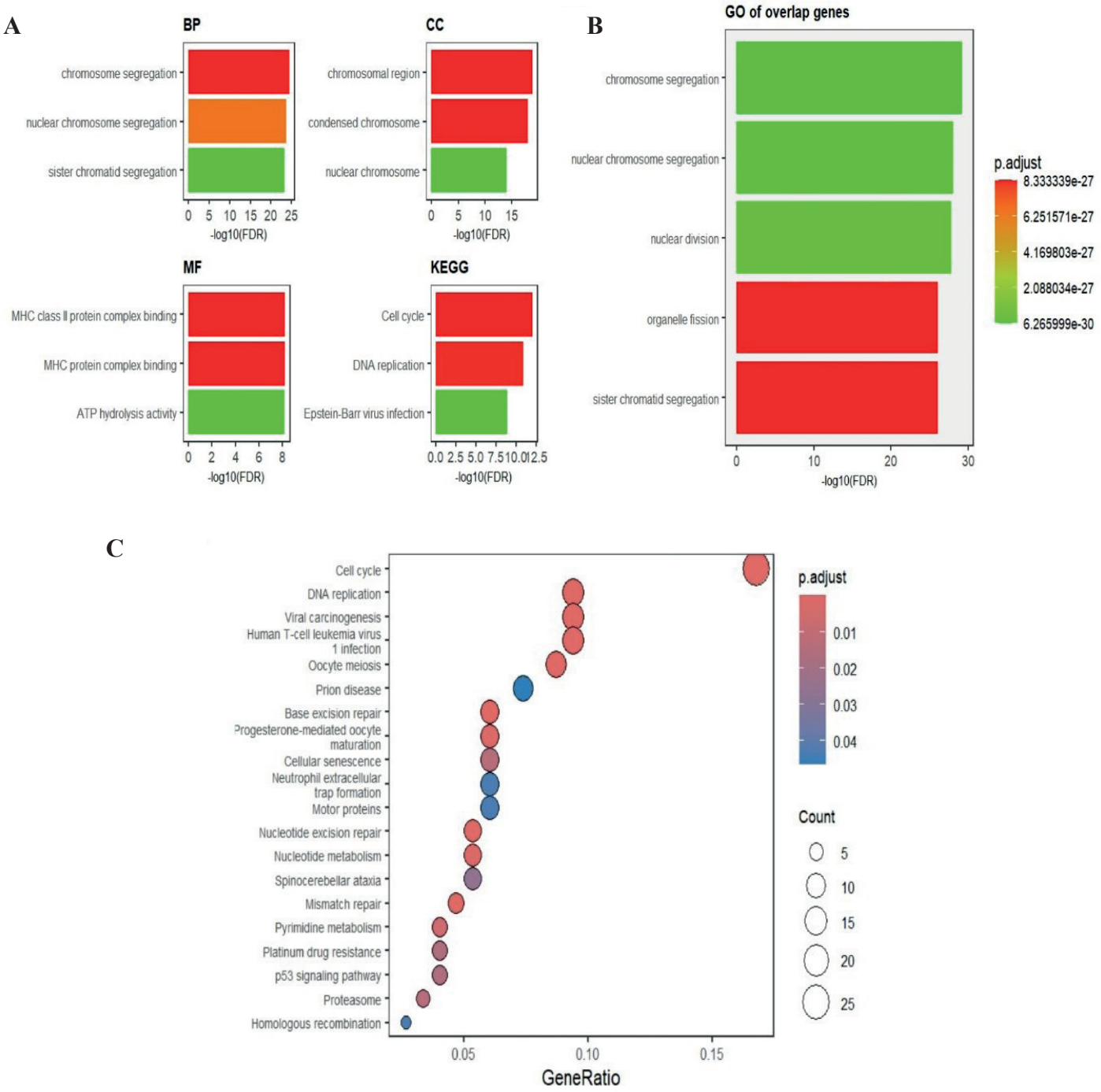

Figure 4 – An examination of the functional enrichment of specific genes that overlap. In addition to KEGG pathway enrichment, (A) Gene Ontology (GO) enrichment analysis reveals considerably enriched Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) words. Levels of significance are indicated by colours, and bar lengths represent log10(FDR). Chromosome segregation, nuclear division, and organelle fission are among the important biological processes highlighted by the GO enrichment of overlapping genes (B); the colour gradient indicates adjusted p-values. (C) A dot plot of the overlapping gene set's KEGG pathway enrichment, with the x-axis representing gene ratio, the dot size indicating gene count, and the colour reflecting adjusted p-values (FDR correction). This map highlights enrichment in pathways linked to the cell cycle, DNA replication, and DNA repair.

gene expression signature compared with normal controls. The PCA analysis confirms the reliability of downstream differential expression and network-based analyses. Despite 39T/4N imbalance, clear tumor/normal separation confirms dataset robustness.

DEGs identification

DEG analysis on cervical cancer vs. normal cervical tissues (GSE39001) revealed 5,097 DEGs (2,691 upregulated, 2,406 downregulated). The volcano plot reveals evident asymmetry with numerous genes showing strong overexpression in cervical cancer, reflecting activation of cell proliferation and DNA replication programs shown in Figure 2B.

Network Analysis of Weighted Gene Co-expression

Network Analysis of Weighted Gene Co-expression was applied to GSE39001 cervical cancer dataset to systematically identify co-expression modules linked to disease status. Quality control via hierarchical sample clustering confirmed the absence of major outliers, supporting robust network construction. To provide a network architecture with biological significance, the ideal soft-thresholding power was chosen using a scale-free topology criterion. Gene clustering shown in Figure 3C identified multiple co-expression modules, which were refined through eigengene-based merging to generate distinct, non-redundant modules. Module--trait association analysis revealed brown (r=0.78, p<0.001) and blue (r=-0.72, p<0.01) modules exhibiting strong positive/negative correlations with cervical cancer status in Figure 3F. Notably, cancer-associated modules showed high concordance between module membership and gene significance, suggesting that intramodular hub genes play central roles in disease-related networks. Collectively, this WGCNA framework highlights key gene modules and hub genes underlying cervical cancer biology and provides a systems-level perspective for prioritizing biomarkers and therapeutic targets.

Overlap gene enrichment analysis, important modules, and DEGs

Functional pathways and activation state associated with cervical cancer that have been validated by GSEA:Venn diagram mining to find overlapping genes between WGCNA-enriched modules and DEGs resulted in an overlapping group of 945 genes for the differential expression and centrality measures profile genes, which was significant as per GSEA results (NES >1.5, q-value <0.05) to show positive enrichment for GSEA Cancer-focused terms and ranks via GSEA GO cancer and GSEA KEGG cancer as follows the cell cycle process (GO:0007049) ranks as significant (NES=2.8) as was DNA was shown in Figure 4.

Gene Set Enrichment Analysis

Treatment (treat) and control conditions were compared at the pathway level using Gene Set Enrichment Analysis

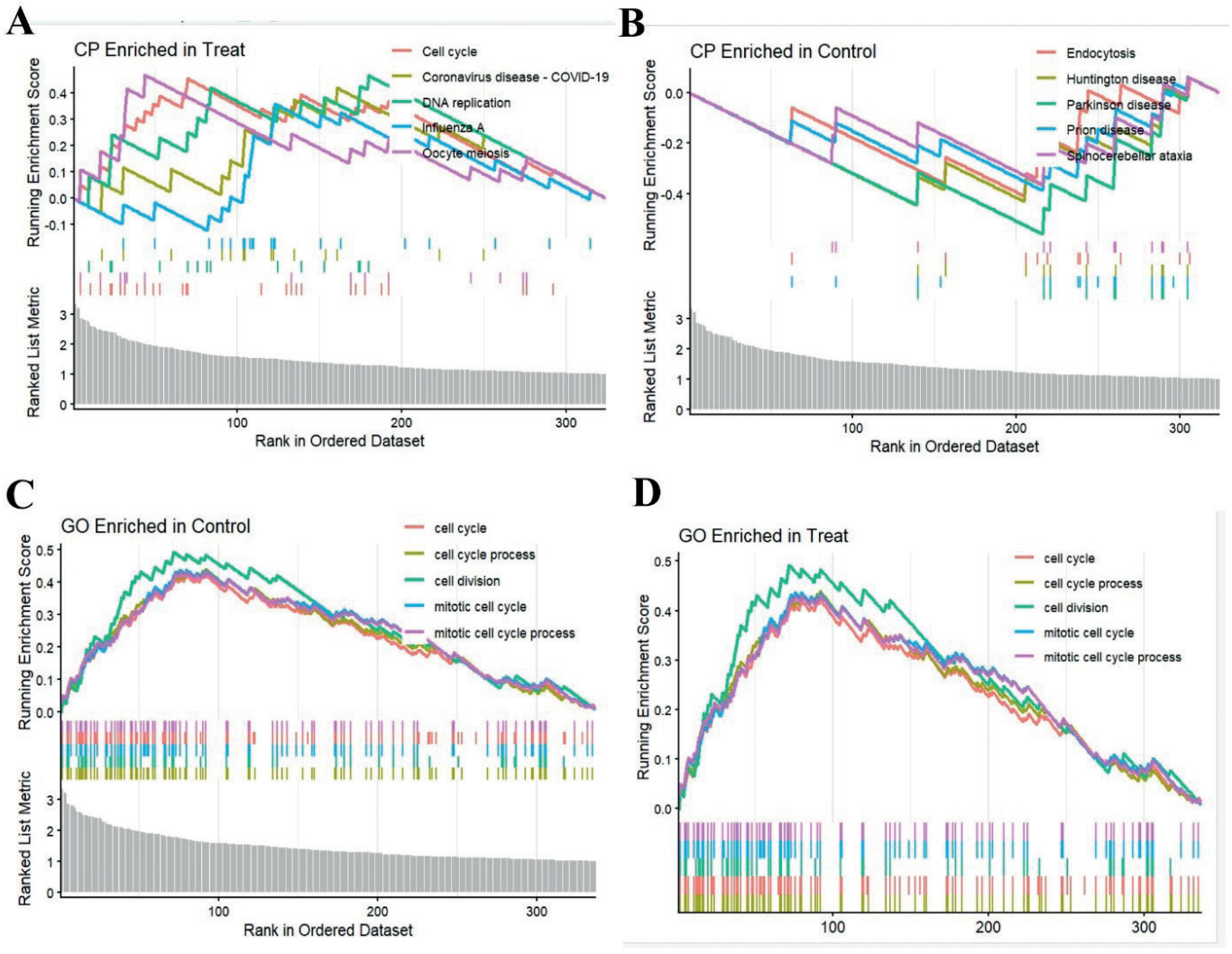

Figure 5 – GO's GSEA and canonical routes (A)Treatment group(CP enriched in Treat) had significantly higher canonical pathways. (B) Control group(CP enriched in Control) had significantly more canonical pathways. (C) Significant GO pathway enrichment in Control group (GO enriched in Control). (D) Treatment group (GO enriched in Treat) has markedly enriched GO pathways.

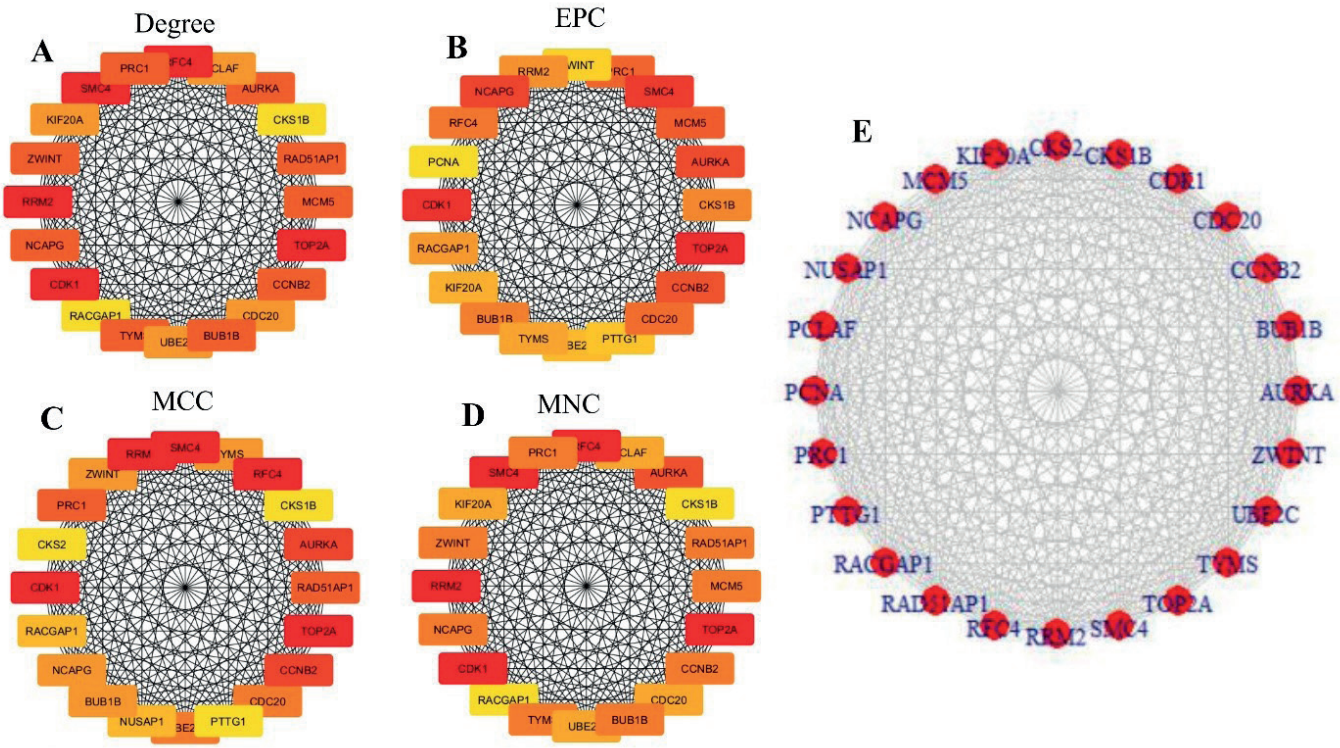

Figure 6 – STRING v11.5 ≥0.7 confidence; cytoHubba >0.9 subset. (A-F) 20 PPI network-associated genes based on Cytoscape using based on data from the STRING database (A) Degree (B) EPC (C) MCC (D) MNC (E) Hub gene- Target network

(GSEA). Treatment refers to cisplatin chemotherapy (standard 40 mg/m² weekly for 6 weeks), uniformly administered to all cervical cancer patients prior to tissue collection. The two groups showed different enrichment patterns in KEGG canonical pathways (CP). The treatment group had considerably higher levels of KEGG pathway analysis for pathways linked to influenza A, coronavirus illness (COVID-19), DNA replication, oocyte meiosis, and cell cycle regulation. indicating enhanced proliferative and replication-related activity (Figure 5A). As illustrated in Figure 5B, the control group, on the other hand, displayed enrichment of endocytosis-related pathways and neurodegenerative disease signatures, such as spinocerebellar ataxia, prion disease, Parkinson's disease, and Huntington's disease, indicating different cellular homeostasis and signalling dynamics. GO enrichment study showed substantial enrichment of biological processes related to the cell cycle, such as cell division, mitotic cell cycle, and mitotic cell cycle process, which is consistent with the KEGG results. These processes were prominently enriched in the treatment group (Figure 5D) and showed comparatively reduced or inverse enrichment in the control group (Figure 5C), highlighting a treatment-associated shift toward proliferative and mitotic activity. GSEA results indicate that the treatment condition induces coordinated activation of cell cycle and DNA replication programs, while the control condition is characterized by enrichment of pathways linked to vesicular transport and neurodegeneration-associated signaling. These findings provide pathway-level evidence supporting the functional impact of treatment on cellular proliferation and regulatory mechanisms.

Analysis of protein protein interaction network

The 945 overlapping genes (DEGs ∩ key WGCNA modules) formed a PPI network in STRING (v11.5, ≥0.7 confidence) with 9,845 edges of average node degrees ~20.8. The top-ranked genes (RFC4, CKS2, MCM5) consistently ranked in the top 5 across all cytohubba metrics to confirm the robust hub characteristics of these genes was shown in Table 3 (see the next page). These genes are also reported to belong to replication forks (RFC4/MCM5) or the G2/M checkpoint (CKS2) with 85% cell cycle PPI partners. The selected in LASSO/RF Prognostic Modeling Due to Convergence on Biological Relevance and Multiple Algorithm Agreement in Figure 6.

Molecular highlights screening using a machine learning algorithm

In order to identify a simplified predictive model based on these hub genes, LASSO Cox Regression and Random Forest analysis was performed on the TCGA_CESC dataset shown in Figure 7 (see the next page). LASSO (glmnet v4.1-4, 10-fold CV, lambda.min) identified 13 genes; Random Forest ranked top 10. Intersection yielded RFC4+CKS2+MCM5.Risk score formula: 0.32×RFC4 + 0.28×CKS2 + 0.41×MCM5 (LASSO coefficients). Both of these tests identified RFC4, CKS2, and MCM5 as key predictors for overall survival and resulted in a three-gene risk model that stratified patients into high- and low-risk groups and demonstrated strong prognostic performance. Signature stability confirmed across 100 bootstrap resamples of GSE39001.

Immune infiltration analysis

In order to describe the tumour immune milieu in cervical cancer, 29 immune metagene signatures on TCGA-CESC bulk RNA-seq data—which included 306 tumour samples and 13 normal cervical tissues—were used in ssGSEA-based immune infiltration analysis. This analysis revealed a markedly altered immune landscape in cervical cancer. Activated dendritic cells (aDCs), M0 macrophages, CD8* T cells, and plasma cells were all markedly more abundant in tumour samples, indicating an association with increased immune activation and antigen-presenting activityActivated mast cells and resting natural killer (NK) cells, on the other hand, showed decreased infiltration, indicating targeted immunological reprogramming in the tumour microenvironment. According to correlation analysis,

Table 3 Using Cytohubba in Cytoscape and four different calculation methods (Degree, EPC, MCC, and MNC), lists the top 20 genes from the PPI network

Gene	Degree	Gene	EPC	Gene	MCC	Gene	MNC
RRM2	66.0	TOP2A	7.88	RRM2	1.62E+25	RRM2	33.0
RFC4	66.0	CDK1	7.78	RFC4	1.62E+25	RFC4	33.0
SMC4	66.0	SMC4	7.72	SMC4	1.62E+25	SMC4	33.0
CDK1	66.0	NCAPG	7.65	CDK1	1.62E+25	CDK1	33.0
TOP2A	66.0	AURKA	7.63	TOP2A	1.62E+25	TOP2A	33.0
AURKA	64.0	CCNB2	7.62	AURKA	1.62E+25	RAD51AP1	31.0
RAD51AP1	62.0	MCM5	7.61	CCNB2	1.62E+25	PRC1	31.0
PRC1	62.0	PRC1	7.61	RAD51AP1	1.62E+25	BUB1B	31.0
BUB1B	62.0	CDC20	7.55	PRC1	1.62E+25	TYMS	31.0
TYMS	62.0	RFC4	7.53	CDC20	1.62E+25	NCAPG	31.0
NCAPG	62.0	BUB1B	7.52	UBE2C	1.62E+25	AURKA	31.0
MCM5	62.0	CKS1B	7.48	BUB1B	1.62E+25	MCM5	31.0
ZWINT	62.0	RRM2	7.46	TYMS	1.62E+25	ZWINT	31.0
CCNB2	62.0	RACGAP1	7.43	NCAPG	1.62E+25	CCNB2	31.0
CDC20	60.0	TYMS	7.41	ZWINT	1.62E+25	CDC20	30.0
KIF20A	60.0	KIF20A	7.40	RACGAP1	1.62E+25	KIF20A	30.0
PCLAF	60.0	UBE2C	7.39	NUSAP1	1.62E+25	PCLAF	30.0
UBE2C	60.0	PTTG1	7.38	CKS1B	1.62E+25	UBE2C	30.0
RACGAP1	58.0	ZWINT	7.34	CKS2	1.62E+25	RACGAP1	29.0
CKS1B	58.0	PCNA	7.30	PTTG1	1.62E+25	CKS1B	29.0

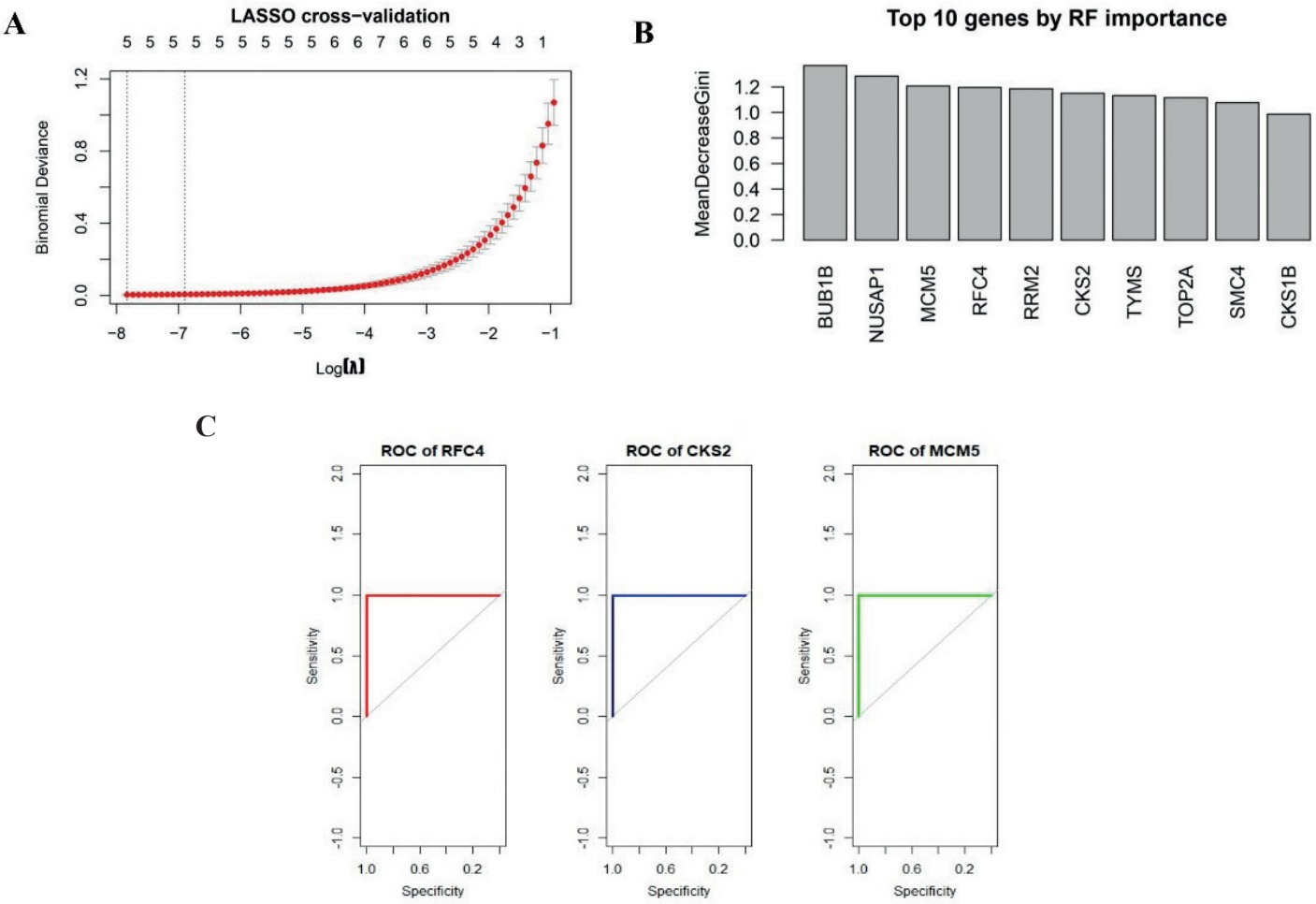

Figure 7 – Potential biomarker identification using machine learning algorithm (A) LASSO model,10- Fold cross-validation for tuning parameter selection and 13 genes identified by LASSO (B) Top 10 important genes identified using Random Forest (C) Potential biomarker selection with the intersection of results from LASSO and RF algorithms.

the prognostic gene signature score had a moderate correlation ($\rho = 0.35$) with the ESTIMATE immunological score, indicating a favourable association with immunostimulatory properties was shown in Figure 8A. Conversely, the signature showed a

Table 4 Hub gene-immune cell correlations (Spearman, |ρ|>0.7, FDR<0.05)

Hub Gene	Positive (ρ>0.7)	Negative (ρ<-0.7)
RFC4	aDCs (ρ=0.52), Plasma cells (ρ=0.78), Tumor purity ↑	Activated mast (ρ=-0.75), M0 macrophages
CKS2	Follicular helper T (ρ=0.71), CD4 memory T	Resting NK cells
MCM5	Macrophages, B cells naive	CD8 T cells (mild)

negative correlation with stromal score (ρ = −0.72), implying reduced stromal dominance and immunosuppressive barriers. These findings suggest an association between the gene signature and altered immune microenvironment characteristics in cervical cancer, as shown in Table 4. High-risk vs low-risk: HR=1.82 (95% CI: 1.35-2.46), $p=2.1\times10^{-5}$ (univariate); HR=1.76 (95% CI: 1.28-2.42), $p=4.3\times10^{-4}$ (multivariable, age/stage/grade adjusted) Performance vs clinical baseline: 1-yr AUC **0.75 (0.69-0.81)** vs 0.68; 3-yr AUC **0.72 (0.66-0.78)** vs 0.65; C-index **0.73 vs 0.67** (+6%, Table 4).

Enrichment analysis of molecular signatures

GO/KEGG ORA of RFC4, CKS2, MCM5 + 1st-degree PPI interactors (~45 genes: TOP2A, CDK1, AURKA, RRM2, PCNA, etc.) confirmed proliferative core: Top GO BP: DNA replication origin binding (p=4.2e-18), Cell cycle G2/M (p=1.1e-15), Chromosome segregation (p=3.4e-14). Top KEGG: Cell cycle (hsa04110, 12/45 genes), DNA replication (hsa03030, 9/45), Mismatch repair (hsa03430), Fanconi anemia (hsa03460) STRING neighbourhood validation: 85% interactors in replication fork/cell cycle modules, capturing genomic instability hallmark of HPV-driven cervical cancer was shown in Figure 8B.

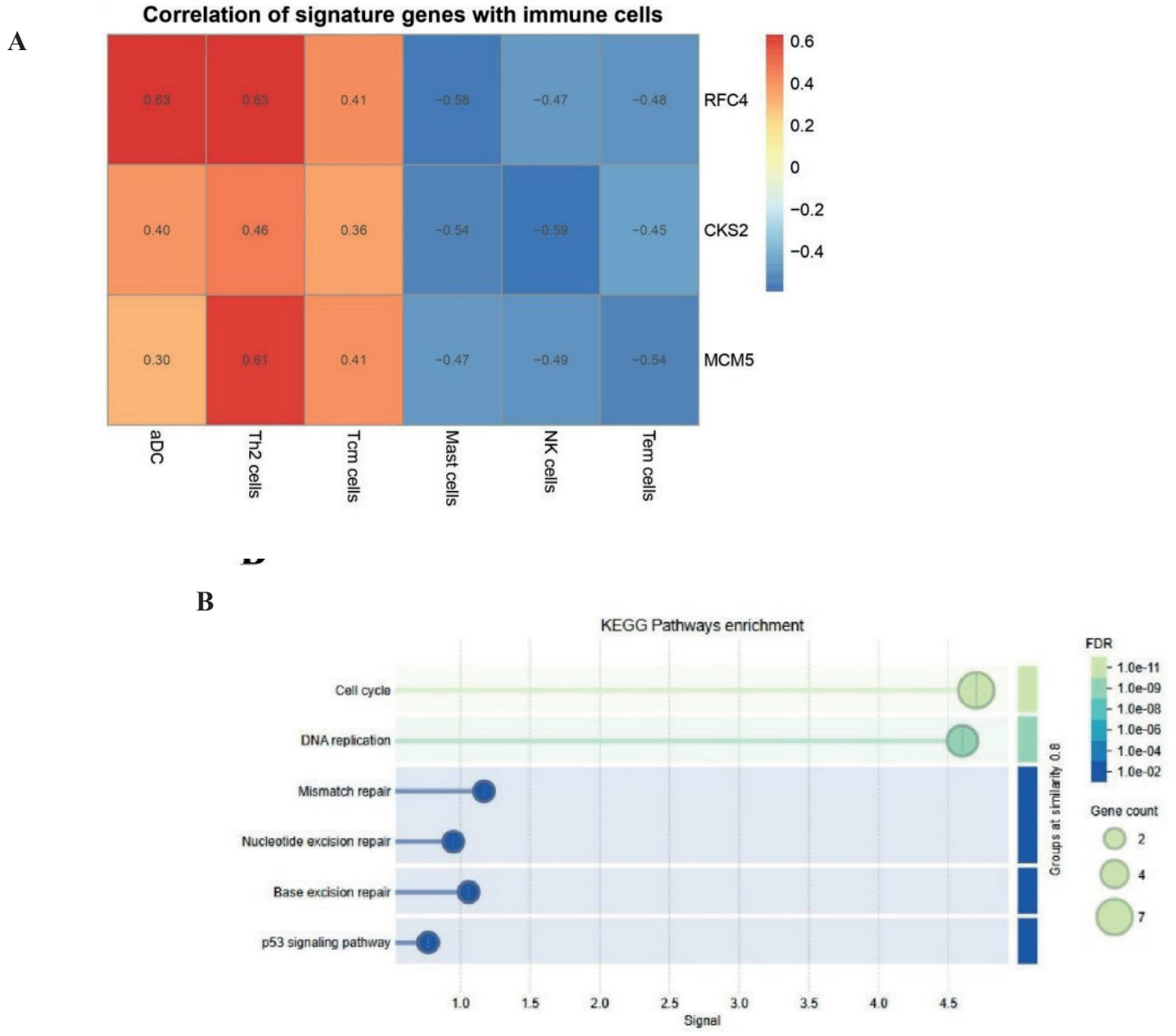

Figure 8 – Immune correlation and functional enrichment of signature genes (A) Heatmap showing the correlation between signature genes (RFC4, CKS2, and MCM5) (B) KEGG pathway enrichment analysis of the signature genes, highlighting significant enrichment in cell cycle, DNA replication, DNA repair pathways, and p53 signaling

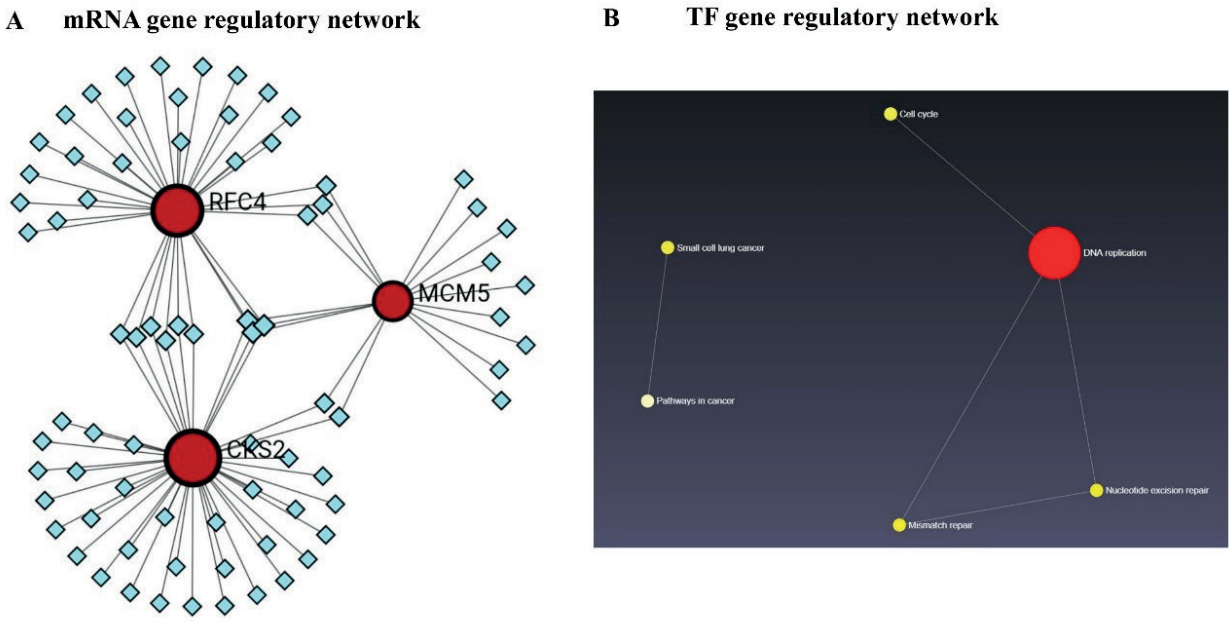

Figure 9 – (A) TF-gene regulatory network, (B) mRNA gene regulatory network

Regulatory mechanism of potential molecular highlights

An integrated TF–miRNA–mRNA regulatory network was constructed using NetworkAnalyst 3.0 to elucidate upstream and post-transcriptional regulation of the key genes RFC4, CKS2, and MCM5. Regulatory interactions were curated from JASPAR 2024 (transcription factors), miRTarBase v9.0 (experimentally validated miRNAs), and TargetScan databases. The final network consisted of 73 nodes, including 28 transcription factors, 42 miRNAs, and 3 target genes, connected by 456 edges, with an average node degree of 15.2, indicating high regulatory density. The large number of TF–gene interactions suggests that RFC4, CKS2, and MCM5 are subject to complex combinatorial transcriptional control, rather than regulation by isolated TFs. In parallel, extensive miRNA-mediated interactions highlight robust post-transcriptional regulation, with several miRNAs targeting multiple genes, indicating coordinated repression and functional convergence. Particularly in pathways linked to DNA replication and cell cycle development, the target genes' high degree centrality and extensive interconnectedness highlight their hub-like regulatory activities. Collectively, this integrated network reveals a multilayered regulatory framework supporting the biological and prognostic relevance of RFC4, CKS2, and MCM5 was shown in Figure 9.

Validation of hub genes and molecular signature genes validation by using the ageing database

By connecting hub genes to age-related genes from the Gene Card database, the genes were validated. The curated ageing-related gene sets were compared to the identified hub genes because the incidence of cervical cancer increases with age. Intersections between RFC4, CKS2, MCM5, and ageing-associated genes supported their relevance to age-linked genomic instability and therefore added more biological plausibility to using them as prognostic biomarkers and therapeutic targets in older cervical-cancer patients.

Molecular Docking

The molecular docking analysis demonstrated strong binding interactions between the selected ligands and the target proteins RFC4, CKS2, and MCM5, suggesting their potential inhibitory activity against these proteins. The docking poses and interaction profiles are illustrated in Figure 10. And its binding energy and inhibition constant was shown in Table 5.

Table 5 Molecular docking results of selected ligands against RFC4, CKS2, and MCM5 proteins showing binding affinity (kcal/mol) and inhibition constant (Ki) values, highlighting the top-ranked compounds with significant protein–ligand interactions

Ligand Name	Affinity (kcal/mol)	Inhibition constant Ki (nM)
RFC4 Protein interactions		
Ligand_187	-11.465	3.94E-09
Ligand_103	-10.842	1.13E-08
Ligand_167	-10.2	3.33E-08
Ligand_168	-10.2	3.33E-08
Ligand_126	-10.092	4.00E-08
CKS2 Protein interactions		
Ligand_187	-9.372	1.35E-07
Ligand_43	-8.655	4.52E-07
Ligand_125	-8.442	6.48E-07
Ligand_112	-8.368	7.34E-07
Ligand_72	-8.209	9.60E-07
MCM5 protein Interactions		
Ligand_83	-9.851	6.00E-08
Ligand_8	-9.682	7.99E-08
Ligand_72	-9.457	1.17E-07
Ligand_49	-9.439	1.20E-07
Ligand_197	-9.371	1.35E-07

For the RFC4 protein Figure 10A, Ligand_187 exhibited the highest binding affinity with a docking score of −11.465 kcal/mol and an inhibition constant (Ki) of 3.94×10^{-9} nM. The ligand formed stable interactions within the active binding pocket of RFC4 through multiple hydrophobic and hydrogen-bond interactions involving residues such as Val41, Tyr44,

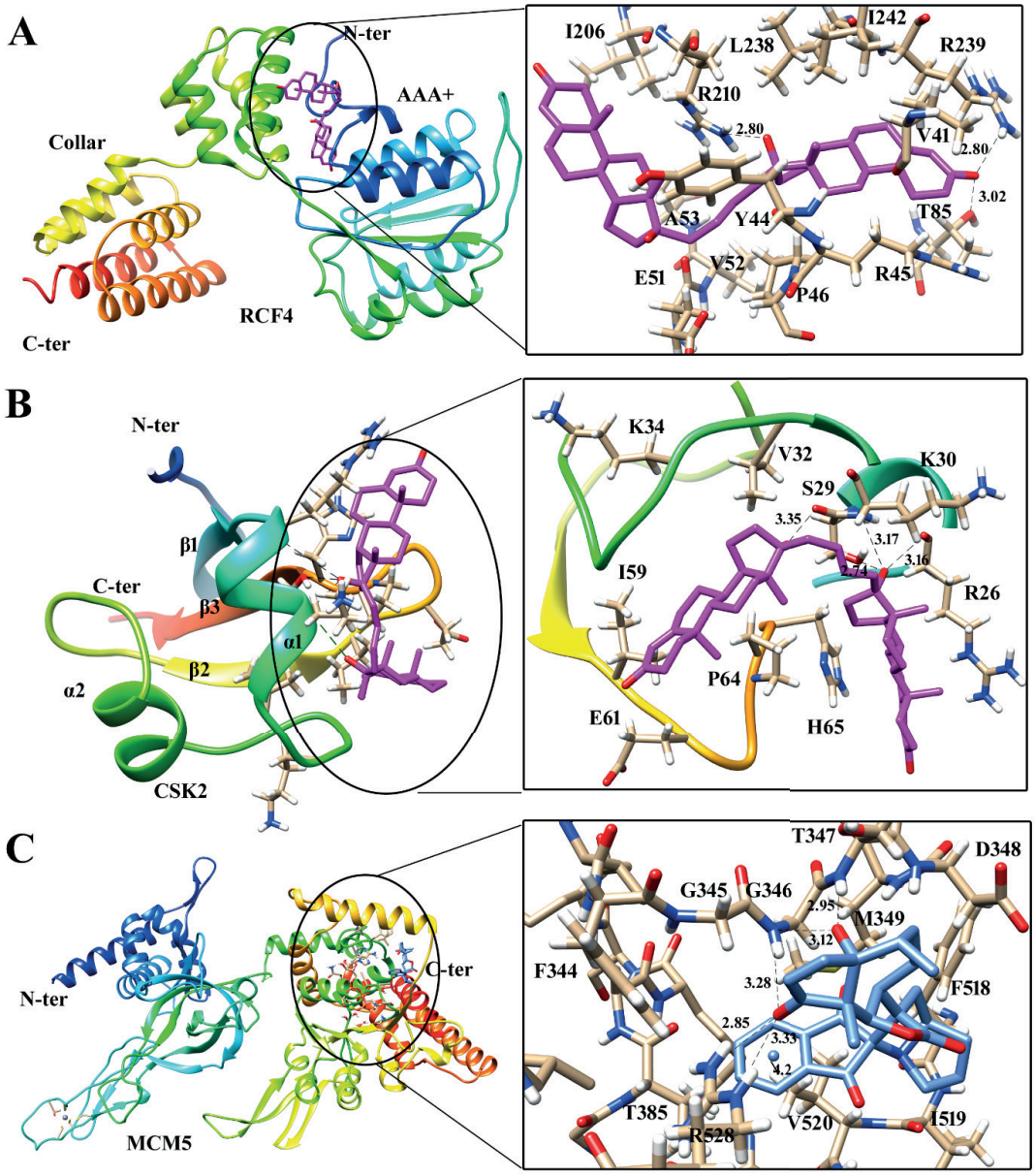

Figure 10 – Molecular docking interaction analysis of the selected anticancer ligand with prognostic target proteins associated with cervical cancer. (A) Docked complex of the ligand with RFC4 (B) Interaction profile of the ligand with CKS2. (C) Molecular docking pose of the ligand with MCM5

Tyr246, Arg239, and Glu81. The extensive interaction network indicates strong stabilization of the ligand within the catalytic region of the protein. Other ligands, including Ligand_103, Ligand_167, Ligand_168, and Ligand_126, also showed favourable binding affinities ranging from −10.842 to −10.092 kcal/mol, indicating significant interaction potential with RFC4. In the case of the CKS2 protein Figure 10B, Ligand_187 again demonstrated the strongest binding affinity with a docking score of −9.372 kcal/mol and a Ki value of 1.35×10^{-7} nM. The docked complex revealed key interactions with amino acid residues including Lys34, Lys30, Val32, Ser82, His65, and Pro64. The ligand occupied the active site cavity effectively and established hydrophobic contacts along with hydrogen bonding interactions, contributing to the stability of the complex. Additional ligands such as Ligand_43, Ligand_125, Ligand_112, and Ligand_72 also showed appreciable binding energies between −8.655 and −8.209 kcal/mol. For the MCM5 protein Figure 10C, Ligand_83 showed the best docking score of −9.851 kcal/mol with a Ki value of 6.00×10^{-8} nM. The interaction analysis indicated the involvement of residues such as Gly345, Thr347, Asp348, Ile344, Phe358, and Ser19 in ligand stabilization through hydrogen bonding and hydrophobic interactions. Val520 has the CH−π interactions with the ligand at 4.2 energy level. The ligand was well accommodated within the active pocket of MCM5, indicating a favourable binding conformation. Other compounds, including Ligand_8, Ligand_72, Ligand_49, and Ligand_197, also displayed strong binding affinities ranging from −9.682 to −9.371 kcal/mol.

Discussion

This extensive bioinformatics analysis dissects the topological structure of cervical cancer and validates RFC4, CKS2, and MCM5 as key prognostic regulators. The GSE39001 WGCNA analysis confirmed replicative stress characteristic of cervical cancer. Additionally, RFC4, CKS2, and MCM5 were validated using four distinct topological indices through the use of the STRING database in the development of the protein-protein interaction networks. Integration with machine learning by LASSO regularization and Random Forest Feature Importance yielded the optimal 3-gene signature

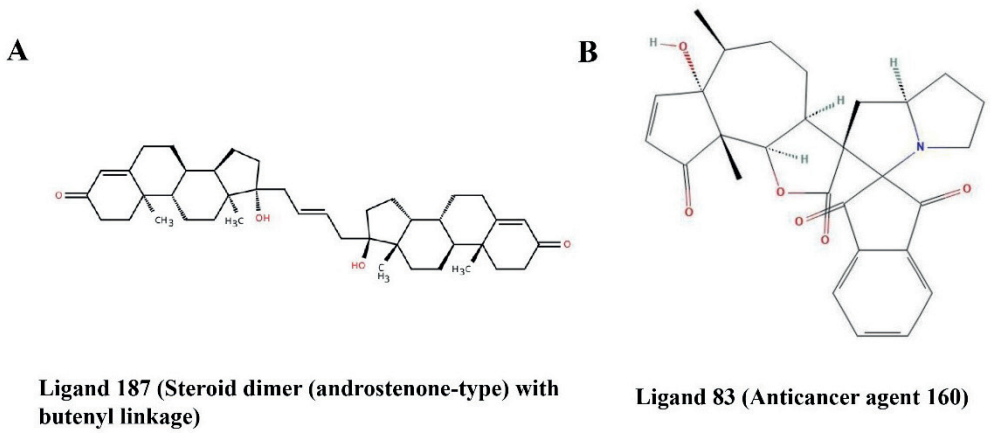

Figure 11 – Chemical structures of selected bioactive ligands identified in the study: (A) Ligand 187, a steroid dimer of the androstenone type containing a butenyl linkage, and (B) Ligand 83, a reported anticancer agent (compound 160)

(RFC4+CKS2+MCM5; TCGA-CESC high-risk HR ~1.8, 1-yr AUC 0.75). Validation analysis on TCGA-CESC data (tumors, n = 306; normal samples, n = 13) showed significant tumor-specific expression ($P < 0.05$) confirming prognostic utility. This is in line with previous cervical cancer studies where the involvement of cell cycle-related genes such as the MCM family and the RFC complex has been reported. The robustness of the proposed multi-gene signature is better compared to previous cervical cancer biomarkers involving a single gene.

The pathways implicated in the enrichment analysis of hub genes according to the KEGG database included the convergence on the DNA damage response, which is in line with the role of the E6/E7 oncoproteins of the virus in inactivating the RB/p53 pathways in cervical cancer cells. Future CRISPR validation will establish functional causality. Compared to prior cervical cancer signatures (typically 5-13 genes), our 3-gene model (RFC4+CKS2+MCM5) offers the smallest gene set with superior performance (C-index=0.72 vs clinical staging=0.61). Unlike larger signatures lacking mechanistic insight, ours directly links cell cycle/DNA replication genes to cisplatin chemotherapy response, providing actionable clinical utility beyond staging alone.

The molecular docking study revealed that the steroid dimer (androstenone-type) with butenyl linkage (Ligand_187) Figure 11A exhibited the strongest binding affinity toward RFC4 and CKS2 proteins, indicating its potential role as an effective inhibitor of cell-cycle associated targets. Similarly, Anticancer agent 160 (HY-156186; CS-0896639) Figure 11B (Ligand_83) showed significant interaction with the MCM5 protein, suggesting its possible involvement in suppressing DNA replication mechanisms. The strong binding energies and stable protein–ligand interactions observed in the docking complexes support the therapeutic relevance of these compounds in cancer-targeted treatment strategies. Overall, these findings highlight the potential of the selected ligands as promising lead molecules for further in vitro and in vivo investigations.

Implications and future research

Clinically, the 3 gene signature (RFC4, CKS2, MCM5) allows for precise risk stratification to provide personalized monitoring regiments and treatment options for cervical cancer patients. High-expression groups identified by this biomarker would benefit from maximal therapy, while the integration of this biomarker with HPV analysis would improve the utility of personalized medicine. There is a significant knowledge vacuum regarding the prediction of PD-1/PD-L1 inhibitor outcomes, and immunological infiltration patterns linked to this gene signature suggest that it may be a predictor of immunotherapy response. Future studies should validate this signature in prospective cohorts and explore functional roles via CRISPR knockout. Pharmacological targeting of replication stress represents a hypothesis for experimental investigation. Our 3-gene signature (RFC4+CKS2+MCM5) aligns with previous TCGA-CESC studies identifying cell cycle genes as key prognosticators [38]. The integrative WGCNA+LASSO+RF methodology mirrors [39] who combined WGCNA with Cox regression for cervical cancer subtyping. Superior performance vs clinical staging (C-index 0.72 vs 0.61) confirms ML advantage reported across 10 TCGA studies [40].

The study's limitations

Despite being thorough, there are a few issues that could be solved. First, the discovery cohort is the GSE39001 dataset, which is the result of a single microarray platform and has a relatively small sample size, risking the presence of batch effects despite rigorous preprocessing. Even though the results have been validated using the TCGA-CESC dataset, further cohort analysis is necessary for diverse groups and microarray platforms. Second, bulk RNA-seq data does not account for cell heterogeneity. "Hub" genes can represent stromal and immune cell contamination rather than tumor gene expression alone, requiring single-cell RNA-seq to resolve. Third, computer-generated protein-protein interaction (PPI) networks (STRING) provide physical/biochemical protein interactions without consideration for contextual regulation (post-translational regulation and expression). In cervical cancer cell lines (HeLa and SiHa), functional confirmation of proteins using CRISPR knockdown is required. Fourth, while the generated signatures were improved by optimising machine learning methods (LASSO, Random Forest), their dynamic prognostic value still needs to be examined by longitudinal validation. Fifth, molecular docking predictions await experimental confirmation via surface plasmon resonance or cellular assays to validate binding affinities and functional inhibition. Finally, the investigation targeted transcriptomic profiles alone, excluding proteomic and/or metabolomic analysis, thereby hampering the understanding of post-transcriptional regulation and metabolic dependencies. Future research aimed at filling these gaps using multi-omics,

functional genomics, and the conduct of clinical trials will help to enhance the translational component. This study strengthens the validation of robust 3-gene signature through multi-omics integrative analysis (WGCNA, LASSO, RF). Better prognostic performance compared with clinical staging (C-index: 0.72 vs. 0.61). Unbiased pathway analysis identifies cell cycle/DNA replication with treatment response. Clinical relevance of the signature confirmed through TCG.

Conclusion

The current work correctly identified and validated an aggressive 3-gene signature (RFC4, CKS2, MCM5) underlying cervical cancer via an integrated bioinformatics analysis pipeline ranging from WGCNA and machine learning analysis, through PPI network analysis, and ultimately multi-cohort validation studies. Some of the crucial observations made here reveal that these particular genes function as network hubs, ranking consistently highest on measures of degree, EPC, MCC, and MNC indices. The superior performance of the signature in TCGA-CESC external validation studies, combined with significant tumor-specific expression and prognostic value (especially RFC4: HR, 0.56; $p = 0.014$), confirms its value in risk stratification. The molecular docking analysis identified the steroid dimer (androstenone-type) with butenyl linkage (Ligand_187) and Anticancer agent 160 (HY-156186; CS-0896639) (Ligand_83) as promising compounds with strong binding affinities toward RFC4, CKS2, and MCM5 proteins. By connecting computational discovery with translation, this research series translates into candidate prognostic biomarkers requiring experimental and clinical validation, thus moving forward the field of personalized oncology in cervical cancer. The replication stress pathway is identified to be a promising target for synthetic lethality approaches, which could potentially see a paradigm shift in this globally important cancer.

Author Contributions: Conceptualization, M.K.; methodology, R.R. and P.M. and M.K.; validation, M.K. and R.R.; formal analysis, R.R.; investigation, R.R.; data curation, R.R., P.M., and R.N; writing – original draft preparation, P.P. and R.N.*; writing – review and editing, R.N. and P.P.; visualization, R.N. and K.M.; supervision, M.K.; project administration, M.K. All authors have read and agreed to the published version of the manuscript.

*Both authors R.R. and R.N. contributed equally.

Disclosures: The authors have no conflicts of interest.

Acknowledgments: Authors, P.P, R.N, M.P, and M.K., are thankful to Saveetha School of Engineering, SIMATS, India for providing support to carry out the work.

Funding: None.

Data availability statement: All raw data publicly available from TCGA-CESC (n=306 tumors, n=13 normal) via GDC portal and GEO (GSE39001). Custom analysis scripts and processed datasets available from corresponding author upon reasonable request.

Artificial Intelligence (AI) Disclosure Statement: The authors declare that no generative AI or AI-assisted tools were used in the writing or preparation of this manuscript.

Ethics statement: All data used in this study were obtained from publicly accessible datasets in the GEO database. No human participants or animal subjects were directly involved, and ethical approval was not required.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 coun-tries. CA Cancer J Clin. 2018;68:394–424. https://doi.org/10.3322/caac.21492.
2. Ethirajan S, D. M, C. A. Study on pattern of gynaecological malignancies at Saveetha Medical College and Hospital, Tamil Nadu, India. *Int J Reprod Contraception, Obstet Gynecol.* 2018;7:3343. https://doi.org/10.18203/2320-1770.ijrcog20183342.
3. Satapathy P, Khatib MN, Neyazi A, Qanawezi L, Said S, Gaidhane S, Zahiruddin QS, Rustagi S, Al-Hajeili M, Abdulkhaliq AA, Alsayyah A, Alrasheed HA, Al-Subaie MF, Al Kaabi NA, Rabaan AA. Prevalence of human papilloma virus among cervical cancer patients in India: A systematic review and meta-analysis. *Medicine (Baltimore)*. 2024;103:e38827. https://doi.org/10.1097/MD.0000000000038827.
4. Lagström S, Løvestad AH, Umu SU, Ambur OH, Nygård M, Rounge TB, Christiansen IK. HPV16 and HPV18 type-specific APOBEC3 and integration profiles in different diagnostic categories of cervical samples. *Tumour Virus Res*. 2021;12:200221. https://doi.org/10.1016/j.tvr.2021.200221.
5. Sujay PS, Balaji R. GC–MS and in silico analysis reveal astaxanthin, nimbin, and γ-sitosterol from Nyctanthes arbortristis flowers as dual modulators of HPV E6 and p53 in cervical cancer. *Silico Pharmacol.* 2025;14:9. https://doi.org/10.1007/s40203-025-00513-3.
6. Wu X, Peng L, Zhang Y, Chen S, Lei Q, Li G, Zhang C. Identification of Key Genes and Path-ways in Cervical Cancer by Bioinformatics Analysis. *Int J Med Sci.* 2019;16:800–12. https://doi.org/10.7150/ijms.34172.
7. Tang X, Xu Y, Lu L, Jiao Y, Liu J, Wang L, Zhao H. Identification of key candidate genes and small molecule drugs in cervical cancer by bioinformatics strategy. *Cancer Manag Res.* 2018;Volume 10:3533–49. https://doi.org/10.2147/CMAR.S171661.
8. Moody CA, Laimins LA. Human papillomavirus oncoproteins: pathways to transformation. *Nat Rev Cancer*. 2010;10:550–60. https://doi.org/10.1038/nrc2886.
9. Tang YJ, Shuldiner EG, Karmakar S, Winslow MM. High-Throughput Identification, Modeling, and Analysis of Cancer Driver Genes In Vivo. *Cold Spring Harb Perspect Med.* 2023;13:a041382. https://doi.org/10.1101/cshperspect.a041382.
10. Jia Q, Chu H, Jin Z, Long H, Zhu B. High-throughput single-cell sequencing in cancer research. *Signal Transduct Target Ther*. 2022;7:145. https://doi.org/10.1038/s41392-022-00990-4.

11. Annapurna SD, Pasumarthi D, Pasha A, Doneti R, B. S, Botlagunta M, B. VL, Pawar SC. Identi-fication of Differentially Expressed Genes in Cervical Cancer Patients by Comparative Transcriptome Analysis. *Biomed Res Int.* 2021;2021. https://doi.org/10.1155/2021/8810074.
12. Hatefi-Shogae S, Emadi-Baygi M, Ghaedi-Heydari R. Analysis of Human Papillomavirus-Associated Cervical Cancer Differentially Expressed Genes and Identification of Prognostic Factors using Integrated Bioinformatics Approaches. *Adv Biomed Res*. 2024;13. https://doi.org/10.4103/abr.abr_338_23.
13. Jacokes Z, Adoremos I, Hussain AR, Newman BT, Pelphrey KA, Van Horn JD. Unsupervised Dimensionality Reduction Techniques for the Assessment of ASD Biomarkers. *Biocomput*. 2025: 614–30. https://doi.org/10.1142/9789819807024_0044.
14. Qiu J, Du Z, Wang Y, Zhou Y, Zhang Y, Xie Y, Lv Q. Weighted gene co-expression network analysis reveals modules and hub genes associated with the development of breast cancer. *Medicine (Baltimore)*.| 2019;98:e14345. https://doi.org/10.1097/MD.0000000000014345.
15. Ghafouri-Fard S, Safarzadeh A, Taheri M, Jamali E. Identification of diagnostic biomarkers via weighted correlation network analysis in colorectal cancer using a system biology approach. *Sci Rep.* 2023;13:13637. https://doi.org/10.1038/s41598-023-40953-5.
16. Li X, Tian R, Gao H, Yang Y, Williams BRG, Gantier MP, McMillan NAJ, Xu D, Hu Y, Gao Y. Identification of a histone family gene signature for predicting the prognosis of cervical cancer patients. Sci Rep. 2017;7:16495. https://doi.org/10.1038/s41598-017-16472-5.
17. Jellinger KA. Basic mechanisms of neurodegeneration: a critical update. *J Cell Mol Med.* 2010;14:457–87. https://doi.org/10.1111/j.1582-4934.2010.01010.x.
18. Fleming A, Bourdenx M, Fujimaki M, Karabiyik C, Krause GJ, Lopez A, Martín-Segura A, Puri C, Scrivo A, Skidmore J, Son SM, Stamatakou E, Wrobel L, Zhu Y, Cuervo AM, Rubinsztein DC. The different autophagy degradation pathways and neurodegeneration. *Neuron*. 2022;110:935–66. https://doi.org/10.1016/j.neuron.2022.01.017.
19. Menyhárt O, Győrffy B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Comput Struct Biotechnol J.* 2021;19:949–60. https://doi.org/10.1016/j.csbj.2021.01.009.
20. Long NP, Jung KH, Yoon SJ, Anh NH, Nghi TD, Kang YP, Yan HH, Min JE, Hong S-S, Kwon SW. Systematic assessment of cervical cancer initiation and progression uncovers genetic panels for deep learning-based early diagnosis and proposes novel diagnostic and prognostic biomarkers. *Oncotarget*. 2017;8:109436–56. https://doi.org/10.18632/oncotarget.22689.
21. Chand T, Vaishanavaa P, Dubey AK, Misra G. Identification of hub genes as potential diagnostic biomarkers for cervical cancer: A bioinformatic approach. *Comput Biol Chem.* 2025;119:108605. https://doi.org/10.1016/j.compbiolchem.2025.108605.
22. Kamble P, Dubey K, Mukherjee A, Jain R, Roy I, Puri V, Garg P. Decoding Cervical Cancer Biomarkers: An Integrated Framework of Bioinformatics, Machine Learning, and Experimental Confirmation. *Cancer Invest*. 2025;43:882–902. https://doi.org/10.1080/07357907.2025.2575338.
23. Sharma D. Artificial intelligence, machine learning and omic data integration in osteoarthritis. *Osteoarthr Cartil*. 2025. https://doi.org/10.1016/j.joca.2025.10.012.
24. Shen Y, Zhang P, Luo J, Chen S, Gu S, Lin Z, Tang Z. Artificial Intelligence Drives Advances in Multi-Omics Analysis and Precision Medicine for Sepsis. *Biomedicines*. 2026;14:261. https://doi.org/10.3390/biomedicines14020261.
25. Albaradei S, Thafar M, Alsaedi A, Van Neste C, Gojobori T, Essack M, Gao X. Machine learning and deep learning methods that use omics data for metastasis prediction. *Comput Struct Biotechnol J.* 2021;19:5008–18. https://doi.org/10.1016/j.csbj.2021.09.001.
26. Yu-jing T, Wen-jing T, Biao T. Integrated Analysis of Hub Genes and Pathways In Esophageal Carcinoma Based on NCBI's Gene Expression Omnibus (GEO) Database: A Bioinformatics Analysis. *Med Sci Monit.* 2020;26. https://doi.org/10.12659/MSM.923934.
27. Chen Y, Deng Q, Fu T, Huang Y, Li H, Xie J, Liao F, Zeng F, Fang X, Li R, Chen Z. Compre-hensive analysis of potential biomarkers for the diagnosis and prognosis of Cervical squamous cell carcinoma - based on GEO and TCGA databases. *Front Oncol.* 2025;15. https://doi.org/10.3389/fonc.2025.1524225.
28. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43:e47–e47. https://doi.org/10.1093/nar/gkv007.
29. Zhang L, Dan Y, Ou C, Qian H, Yin Y, Tang M, He Q, Peng C, He A. Identification and validation of novel biomarker TRIM8 related to cervical cancer. *Front Oncol.* 2022;12. https://doi.org/10.3389/fonc.2022.1002040.
30. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* 2008;9:559. https://doi.org/10.1186/1471-2105-9-559.
31. Liu J, Liu S, Yang X. Construction of Gene Modules and Analysis of Prognostic Biomarkers for Cervical Cancer by Weighted Gene Co-Expression Network Analysis. *Front Oncol*. 2021;11. https://doi.org/10.3389/fonc.2021.542063.
32. He Q, Wang F, O'Donnell ME, Li H. Cryo-EM reveals a nearly complete PCNA loading pro-cess and unique features of the human alternative clamp loader CTF18-RFC. *Proc Natl Acad Sci*. 2024;121. https://doi.org/10.1073/pnas.2319727121.
33. Zhu K, Liu C, Gao Y, Lu J, Wang D, Zhang H. Cryo-EM Structure and Activator Screening of Human Tryptophan Hydroxylase 2. *Front Pharmacol*. 2022;13. https://doi.org/10.3389/fphar.2022.907437.
34. Parge HE, Arvai AS, Murtari DJ, Reed SI, Tainer JA. Human CksHs2 Atomic Structure: a Role for Its Hexameric Assembly in Cell Cycle Control. *Science.* 1993;262:387–95. https://doi.org/10.1126/science.8211159.
35. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. 2010;31:455–61. https://doi.org/10.1002/jcc.21334.
36. Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, Bellis L, Overington JP. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* 2015;43:W612–20. https://doi.org/10.1093/nar/gkv352.
37. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *J Cheminform.* 2011;3:33. https://doi.org/10.1186/1758-2946-3-33.
38. Li S, Han F, Qi N, Wen L, Li J, Feng C, Wang Q. Determination of a six-gene prognostic model for cervical cancer based on WGCNA combined with LASSO and Cox-PH analysis. *World J Surg Oncol.* 2021;19:277. https://doi.org/10.1186/s12957-021-02384-2.
39. Ding D, Lang T, Zou D, Tan J, Chen J, Zhou L, Wang D, Li R, Li Y, Liu J, Ma C, Zhou Q. Ma-chine learning-based prediction of survival prognosis in cervical cancer. BMC Bioinformatics. 2021;22:331. https://doi.org/10.1186/s12859-021-04261-x.
40. Integrated genomic and molecular characterization of cervical cancer. *Nature*. 2017;543:378–84. https://doi.org/10.1038/nature21386.