International Journal of Women's Health

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Integrated Genomic Analysis Reveals New Diagnostic Biomarkers and Immune Mechanisms for Polycystic Ovary Syndrome

Ning Huang[1], LuYun Lou[2]

[1]Cixi Integrated Traditional Chinese and Western Medicine Healthcare Group, Ningbo, Zhejiang, 315300, People's Republic of China; [2]Henghe Central Health Center of Cixi City, Ningbo, Zhejiang, 315300, People's Republic of China

Correspondence: Ning Huang; LuYun Lou, Email xiaojinyu0229@hotmail.com; huangxm1999@hotmail.com

Objective: Polycystic ovary syndrome (PCOS) is a prevalent endocrine disorder associated with metabolic dysregulation and chronic inflammation. This study employed bioinformatics approaches to analyze the roles of neutrophil extracellular traps (NETs)-related genes (NETRGs) and mitochondria-related genes (MRGs) in PCOS pathogenesis.

Methods: Through differential expression analysis, enrichment studies, and machine learning models, we identified 18 differentially expressed neutrophil extracellular trap- and mitophagy-related genes (NETMRDEGs), such as *S100A9, MYH9*, and *ATG5*. Functional enrichment revealed their involvement in IL-17 signaling and neutrophil chemotaxis. A LASSO regression model incorporating these genes demonstrated high diagnostic accuracy (AUC > 0.9). Immune infiltration analysis highlighted activated B cells and effector memory $CD8^+$ T cells as key contributors. Moreover, protein-protein interaction (PPI) networks further elucidated functional synergies among NETMRDEGs.

Results: The study identified distinct molecular clusters in PCOS patients based on the expression of NETMRDEGs. Cluster 2 exhibited higher immune infiltration (eg, gamma delta T cells and eosinophils) and severe metabolic dysfunction. The LASSO model achieved superior diagnostic performance (AUC: 0.93) compared to traditional biomarkers such as testosterone (AUC: 0.68). Experimental validation in a PCOS mouse model confirmed elevated expression of hub genes (*S100A9, MYH9*, and *ATG5*) and increased granulosa cell apoptosis.

Conclusion: The interaction between NETRGs and MRGs forms a "dual-engine" mechanism driving PCOS pathogenesis. This study proposed a novel diagnostic model and therapeutic targets (eg, DNase I and urolithin A), advancing precision medicine for PCOS.

Keywords: polycystic ovary syndrome, neutrophil extracellular traps, mitochondria-related genes, immune infiltration, diagnostic model

Introduction

Polycystic ovary syndrome (PCOS) is a highly prevalent reproductive endocrine disorder with a multifactorial and complex pathogenesis, encompassing metabolic disturbances, chronic low-grade inflammation, and perturbations in the local ovarian immune microenvironment. Emerging evidence over the past decade has underscored that neutrophil extracellular trap (NET) formation and mitochondrial dysfunction are critical drivers of inflammatory responses and follicular developmental defects.

However, the intricate crosstalk mechanism between these two pathways in the initiation and progression of PCOS remains poorly defined. Herein, based on integrated multi-omics data analysis, this study is designed to explore the synergistic regulatory roles of NETs-related genes and mitophagy-related genes in PCOS, screen for potential molecular biomarkers, and construct a robust diagnostic model. This investigation is expected to shed new light on the immuno-metabolic characteristics of PCOS and provide a theoretical basis for advancing precise diagnosis and targeted therapeutic strategies for this disorder.

https://doi.org/10.2147/IJWH.S560802
Received: 19 August 2025
Accepted: 7 March 2026
Published: 8 April 2026

Materials and Methods

Data Download

Through the R package GEOquery[1] (Version 2.70.0) from a GEO database[2] (https://www.ncbi.nlm.nih.gov/geo/), PCOS datasets GSE34526[3] and GSE137684 were downloaded. Samples were from Homo sapiens (Table 1).

The chip platform of dataset GSE34526 was GPL570, including 7 PCOS samples and 3 control samples. Dataset GSE137684 included 8 PCOS samples and 4 control samples. The tissue sources of these two datasets were (granulosa cells Tissue), and all PCOS and control groups were included. The samples incorporated in these two datasets were both derived from Asian female populations, with GSE34526 consisting of Indian women and GSE137684 consisting of Chinese women. Given that these datasets primarily characterize the molecular landscape of Asian populations, further validation across different ethnic groups and multi-center cohorts is required to confirm the generalizability and robustness of our findings.

NET-related genes (NETRGs) were retrieved based on the GeneCards database[4] (https://www.genecards.org/). Specifically, with the keyword of "Neutrophil Extracellular Traps" and only "protein-coding" NETRGs kept, 155 NETRGs were obtained. Additionally, with "Neutrophil Extracellular Traps" as the search keyword, 146 NETRGs were obtained from the published literature[5,6] in PubMed. Finally, we merged and deduplicated the NETRGs from the GeneCards database and PubMed literature, and 258 NETRGs were finally obtained (Table S1).

Mitophagy-related genes (MRGs) were collected through the GeneCards[4] (https://www.genecards.org/). Specifically, with the keyword of "Mitophagy" and only "protein-coding" MRGs retained, 3429 MRGs were identified. Additionally, 29 MRGs were clarified based on the MSigDB database[7] using "Mitophagy" as a search term. Finally, MRGs from the above two databases were merged to obtain 3429 MRGs (Table S2).

Finally, NETRGs and MRGs were overlapped to obtain 68 NET- and mitophagy-related genes (NETMRGs) (Table S3).

After debatching GSE34526 and GSE137684 datasets utilizing the R package sva[8] (3.50.0), the combined dataset was obtained, including 15 PCOS cases and 7 control cases. Afterward, the R package limma[9] (3.58.1) was employed for standardizing this integrated GEO dataset, followed by the standardization and normalization of annotation probes. Principal component analysis (PCA)[10] was conducted on the expression matrices before and after batch effect removal. As a dimensionality reduction approach, PCA reduces the features of high-dimension data and transforms data into low dimension, along with displaying features in 2D or 3D graphs.

Differentially Expressed NETMRGs (NETMRDEGs) Associated with PCOS

Based on the combined datasets, samples were grouped: PCOS and control. Genes in these two groups underwent differential analysis utilizing the R package limma (3.58.1). With $|\log FC| > 0$ and $p < 0.05$, differentially expressed genes (DEGs) were identified, including unregulated DEGs ($\log FC > 0$, $p < 0.05$) and downregulated DEGs ($\log FC < 0$, $p < 0.05$). Benjamini-Hochberg (BH) was the p-value correction approach. Volcano plots were generated for displaying difference analysis results using the R package ggplot2 (3.4.4).

DEGs with $|\log FC| > 0$ and $p < 0.05$ in combined datasets were intersected with NETMRDEGs to obtain PCOS-related NETMRDEGs, which was visualized using the Venn diagram. The top 20 NETMRDEGs were displayed through a heatmap drawn by the R package pheatmap (1.0.12).

Table 1 GEO Microarray Chip Information

	GSE34526	GSE137684
Species	Homo sapiens	Homo sapiens
Platform	GPL570	GPL17077
Samples in PCOS group	7	8
Samples in Control group	3	4
PMID	PMID:22904171	/

Abbreviations: GEO, Gene Expression Omnibus; PCOS, Polycystic ovary syndrome.

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) Analyses

GO analysis[11] can be used to analyze functional enrichment, involving biological process (BP), molecular function (MF), and cell component (CC). KEGG[12] can provide data about genomes, pathways, and diseases. NETMRDEGs underwent GO and KEGG enrichment analyses utilizing the R package clusterProfiler[13] (4.10.0), with an entry screening standard of $p < 0.05$ and FDR value (q) < 0.05, and BH as the *p*-value correction approach.

Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA)[14] is used to evaluate the distribution trend of genes from a predefined gene set within a list of genes ranked by their correlation with a phenotype, in order to assess their contribution to the phenotype. In this study, genes from the combined GEO datasets were ranked based on logFC values between the PCOS group and the Control group. Subsequently, the R package clusterProfiler (Version 4.10.0) was employed to conduct GSEA on all genes in the combined GEO datasets. The parameters used in this GSEA were as follows: seed set to 2020, with a minimum gene set size of 10 and a maximum size of 500. Gene sets c2.all.v2023.2.Hs.symbols were obtained from the Molecular Signatures Database (MSigDB) for the GSEA. The screening criteria for the gene set enrichment analysis were a p value < 0.05 and an FDR value (q value) < 0.25.

Gene Set Variation Analysis (GSVA)

Gene Set Variation Analysis (GSVA)[15] is a non-parametric, unsupervised analytical method designed to assess the enrichment results of gene sets within a transcriptome microarray by converting the expression matrix of genes across different samples into an expression matrix of gene sets. This approach evaluates whether various pathways are enriched across different samples. The c2.cp.v2023.2.Hs.symbols.gmt gene set was obtained from the Molecular Signatures Database (MSigDB).[16] Using the R package GSVA (Version 1.50.0), we conducted GSVA on the integrated GEO dataset (Combined Datasets) to calculate the functional enrichment differences between the polycystic ovary syndrome (PCOS) group and the control group. The screening criterion for GSVA is a p-value < 0.05.

Model Establishment for PCOS Diagnosis

For generating PCOS diagnostic models from the combined datasets, NETMRDEGs were analyzed by Logistic regression. Independent variables' relationship with dependent variable was analyzed through Logistic regression with a binary variable as the dependent variable (PCOS and control groups). NETMRDEGs were identified with $p < 0.05$, followed by constructing a Logistic regression model.

Next, with seed (500) and family = "binomial",[17] Logistic regression model Absolute Shrinkage and Selection Operator (LASSO) (Least) was performed on NETMRDEGs included based on the linear regression analysis, and the penalty term (lambda × absolute value of slope) was added for reducing model overfitting and improving its generalization. First, genes overlapping between the standardized gene expression matrix and the list of NET & MRDEGs differential genes were extracted to construct a candidate feature set. The first 7 samples were defined as the control group, and the subsequent 15 samples as the polycystic ovary syndrome (PCOS) group, which served as the binary outcome variable. A LASSO regression model was fitted using the glmnet package with the parameters set as family = "binomial" and alpha = 1. Ten-fold cross-validation was performed to determine the optimal penalty parameter λ corresponding to the minimum cross-validation error (λ.min). Genes with non-zero coefficients under this λ.min value were extracted as core feature genes. Diagnostic model diagram and variable trajectory diagram were employed for visualizing LASSO results (POST diagnostic model), and NETMRDEGs included served as the Key Genes. After that, according to the risk coefficient of regression, the calculation of LASSO risk score (RiskScore) was conducted based on the formula below:

$$\text{Risk}Score = \sum_i Coefficient\ (gene_i) * mRNA\ Expression\ (gene_i)$$

https://doi.org/10.2147/IJWH.S560802

Then, based on NETMRDEGs included in the regression model, the support vector machine (SVM)[18] algorithm was employed for SVM model construction. According to the amount of genes whose accuracy was the highest and error rate was the lowest, NETMRDEGs were identified.

NETMRDEGs were selected from LASSO regression analysis and SVM analysis, and the intersection genes were used for later analysis.

Validation of the PCOS Diagnostic Model

Next, a Nomogram was constructed for Key Genes utilizing R package rms, which can display independent variables' association in a rectangular coordinate system through clustered disjoint line segments. Diverse variables in the multiple regression model was characterized utilizing a specific scale, with the total score predicting event occurrence.

The fitting between actual and predicted model probability was drawn in calibration plot for evaluating the model's predictive value.

Decision curve analysis (DCA) can assess predictive models, diagnostic detections, and biomarkers. Finally, the Logistic regression model's accuracy and discrimination were evaluated via DCA plot drawn through R package ggDCA.

In addition, ROC curve plotting was performed utilizing R package pROC, followed by calculation of area under the curve (AUC) value in the combined datasets for evaluating the diagnosis performance of linear predictors of PCOS. The AUC ranged from 0.5 to 1, with its value closer to 1 indicating a better value for diagnosis.

Protein-Protein Interaction (PPI)

PPI network includes interacting individual proteins. Protein (known and predicted) interactions can be searched through the STRING database.[19] Specifically, with the biological species as human, and the confidence level ≥ 0.150, a PPI network was constructed utilizing the STRING database, and then visualized through Cytoscape.

Genes with similar functions were predicted from the selected Key Genes using the GeneMANIA website,[20] followed by constructing an interaction network.

Establishment of Regulatory Network

Transcription factors (TFs) control gene expression through interactions with Key Genes during the post-transcription period. TF was retrieved via the ChIPBase database[21] (http://rna.sysu.edu.cn/chipbase/), which was then analyzed for its regulation on Key Genes; the mRNA-TF network was visualized utilizing Cytoscape[22] software.

Additionally, miRNAs are crucial for biological development, which could modulate various genes, and one gene could also be modulated by diverse miRNAs. Key Genes and their associations with miRNA were analyzed through the TarBase[23] database (http://www.microrna.gr/tarbase). Cytoscape software was applied for the visualization of the mRNA-miRNA network.

RNA-binding protein (RBP)[24] is critical for regulating genes and life activities (RNA synthesis, alternative splicing, ect). Based on StarBase v3.0 database[25] (https://starbase.sysu.edu.cn/), the Key Genes of the target RAP were predicted, and Cytoscape software was employed for visualizing the mRNA-RBP network.

Finally, through the Comparative Toxicogenomics Database[26] (https://ctdbase.org/), drug targets of Key Genes were predicted. Key Gene-drug interactions were explored, followed by visualizing the mRNA-drug Network utilizing the Cytoscape software.

Differential Expression, Correlation, and Receiver Operating Characteristic (ROC) Curve Analyses

Key Genes' expression differences between the PCOS and control groups in the combined datasets were analyzed, and the corresponding maps were plotted. After that, ROC curves were plotted utilizing R package pROC (1.18.5), and the AUC values were calculated as it can assess the diagnostic effect of Key Genes' expression on PCOS occurrence. AUC ranged from 0.5 to 1, and an AUC value closer to 1 indicates a better diagnostic value, with a value of 0.5–0.7, 0.7–0.9, and above 0.9 indicating a low, moderate, and high accuracy, respectively.

Furthermore, correlation analysis was conducted on Key Genes' expression in the combined datasets, and the correlation heatmap was plotted via the R package pheatmap (1.0.12). Specifically, an absolute value of correlation coefficient (r value) < 0.3, 0.3–0.5, 0.5–0.8, and > 0.8 indicates weak/no correlation, weak, moderate, and strong correlation, respectively.

Immune Infiltration Analysis

Single-sample GSEA (ssGSEA)[27] can quantify each immune cell infiltrate's relative abundance. In short, infiltrating immune cell subtypes were first labeled: activated CD8 T cells and dendritic cells, regulatory T cells, etc.

Next, with the enrichment scores obtained from ssGSEA representing each immune cell infiltrate's relative abundance, a corresponding immune cell infiltrate matrix was obtained. After that, the comparison maps were plotted utilizing R package ggplot2 (3.4.4), and immune cells in the integrated GEO dataset between the LASSO RiskScore LowRisk group and HighRisk group were displayed.

Finally, differential immune cells were identified. Moreover, Spearman algorithm was employed for correlation analysis between immune cells, followed by displaying the results utilizing R package pheatmap (1.0.12). Additionally, Key Genes' correlation with immune cells was analyzed using Spearman algorithm; the results were visualized through bubble plots using R package ggplot2 (3.4.4).

Statistical Analysis

R software (4.2.2) was employed for data analysis. Regarding continuous variables between two groups, variables with and without normal distribution were compared utilizing the independent Student's *t*-test and Mann–Whitney *U*-Test (Wilcoxon Rank Sum Test), respectively. Moreover, data among multiple groups were compared using the Kruskal–Wallis test. Additionally, the correlation coefficient was calculated through Spearman analysis. A two-sided *p*-value of < 0.05 indicates statistical significance.

Results

Technology roadmap: (Figure 1)

Data Collection and Correction

Firstly, the R package sva was used to remove batch effect on polycystic ovary syndrome (PCOS) Datasets GSE34526 and GSE137684 to obtain Combined GEO datasets. Subsequently, the distribution boxplots (Figure 2A and B) were used to compare the expression values of the datasets before and after batch effect removal. Secondly, PCA (Principal Component Analysis) plot (Figure 2C and D) was used to compare the distribution of low-dimensional features before and after batch effect removal. The results of distribution boxplot and PCA plot showed that the batch effect of samples in the polycystic ovary syndrome (PCOS) dataset was basically eliminated after batch removal.

Polycystic Ovary Syndrome Related Neutrophil Extracellular Traps and Mitophagy Related Differentially Expressed Genes

To identify Neutrophil Extracellular Traps and mitophagy-related genes (NETMRGs), we identified neutrophil extracellular traps-related genes (neutrophil extracellular traps-related genes, NETRGs) and mitochondrial autophagy related gene (Mitophagy - related genes, MRGs) get intersection and map Wayne (Figure 3A), get 68 neutrophil extracellular trap net and mitochondrial autophagy related gene (NETMRGs).

The data of the Combined GEO Datasets were divided into polycystic ovary syndrome (PCOS) group and Control group, respectively. To analyze the differences in gene expression values between the polycystic ovary syndrome (PCOS) group and Control group in the Combined GEO Datasets, The R package limma was used for differential analysis of the Combined GEO Datasets to obtain differentially expressed genes between the two sets of data, and the results were as follows: A total of 2745 differentially expressed genes (DEGs) met |logFC| > 0 and p value < 0.05 threshold were identified in the Combined Datasets. Under this threshold, there were 1344 up-regulated genes (logFC > 0 and p value < 0.05) and

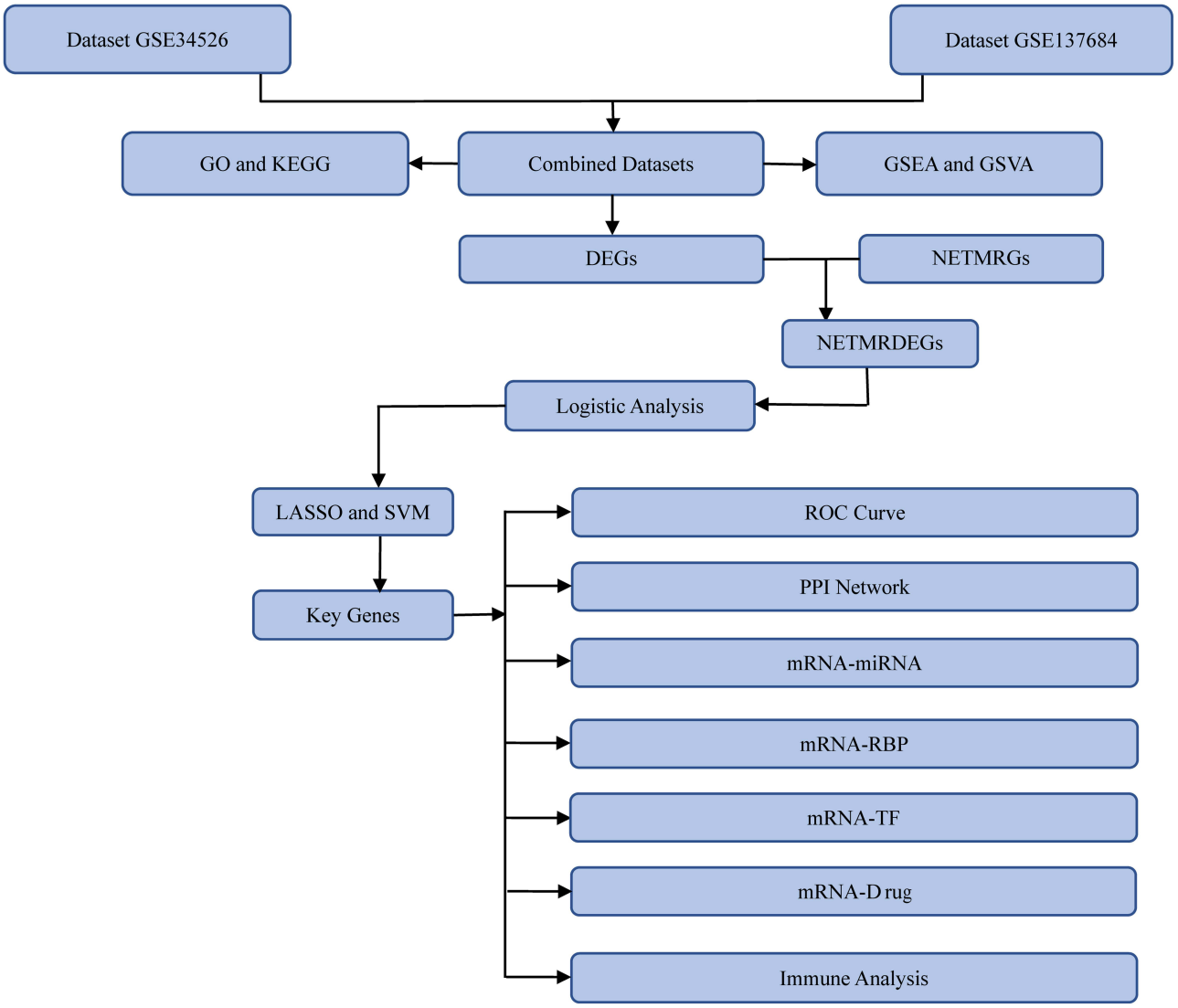

Figure 1 Technology Roadmap.
Abbreviations: GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; GSEA, Gene Set Enrichment Analysis; GSVA, Gene Set Variation Analysis; DEGs, Differentially Expressed Genes; NETMRGs, Neutrophil Extracellular Traps and Mitophagy-Related Genes; NETMRDEGs, Neutrophil Extracellular Traps and Mitophagy-Related Differentially Expressed Genes; LASSO, Least Absolute Shrinkage and Selection Operator; SVM, Support Vector Machine; ROC, Receiver Operating Characteristic; PPI, Protein-Protein Interaction; RBP, RNA-Binding Protein; TF, Transcription Factor.

1401 down-regulated genes (logFC < 0 and p value < 0.05). The volcano map was drawn based on the difference analysis results of this dataset (Figure 3B).

To obtain neutrophil extracellular traps and mitophagy related differentially expressed genes (NETMRDEGs), All differentially expressed genes (DEGs) with |logFC| > 0 and p value < 0.05 and neutrophil extracellular traps and mitophagy-related genes (NETMRGs) were intermixed and Venn diagram was drawn (Figure 3C). A total of 18 neutrophil extracellular trap and mitophagy-related differentially expressed genes (NETMRDEGs) were obtained, which were S100A9, MYH9, SERPINA1, CD44, TREM1, CCL2, RAC2, ABCA1, TKT, S100A8, NLRP3, ATG7, PDK4, MFN2, ATG5, AKT1, IL1B, ACTN4. According to the intersection results, For the integrated GEO dataset (Combined The expression differences of neutrophil extracellular traps and mitophagy-related differentially expressed genes (NETMRDEGs) between different sample groups in Datasets were analyzed, and the R package pheatmap was used to draw a heatmap to show the Top20 neutrophil extracellular traps and mitophagy-related differentially expressed genes (netMRdegs) s) analysis results (Figure 3D).

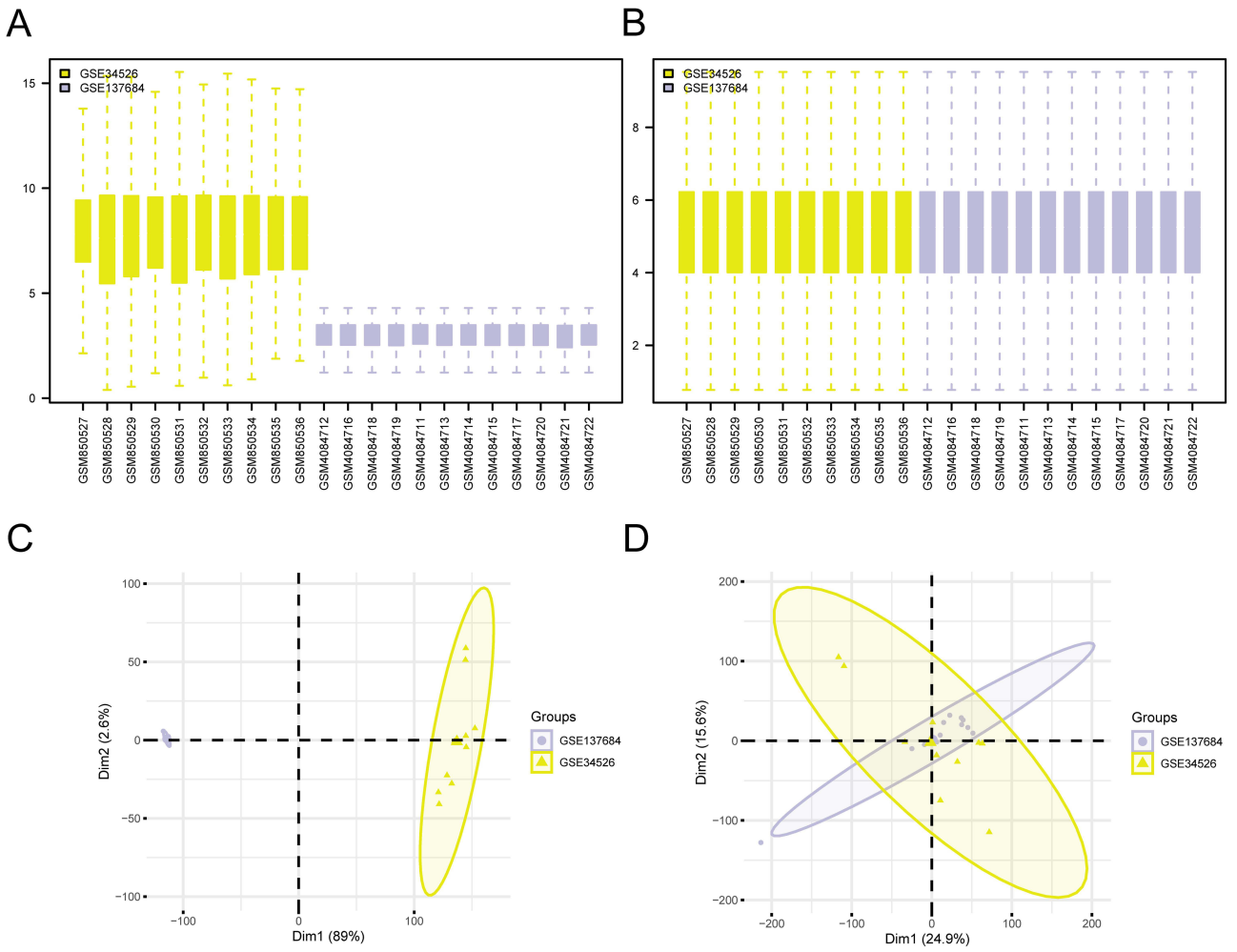

Figure 2 Batch Effects Removal of GSE34526 and GSE137684. (**A**) Box plot of Combined GEO Datasets distribution before batch removal. (**B**) Post-batch integrated GEO Datasets (Combined Datasets) distribution boxplots. (**C**) PCA plot of the datasets before debatching. (**D**) Go to the PCA map of the Combined GEO Datasets after batch processing. Polycystic ovary syndrome (PCOS) dataset GSE34526 is yellow, and polycystic ovary syndrome (PCOS) dataset GSE137684 is purple.
Abbreviations: PCA, Principal Component Analysis; PCOS, Polycystic ovary syndrome.

Gene Ontology (GO) and Pathway (KEGG) Enrichment Analysis

Gene ontology (GO) and pathway (KEGG) enrichment analysis were performed to further explore the biological process (BP), cellular component (CC), and molecular mechanism of 18 neutrophil extracellular traps and mitophagy-related differentially expressed genes (NETMRDEGs). The relationship between molecular function (MF) and biological pathway (KEGG) and polycystic ovary syndrome (PCOS) was explored. The 18 neutrophil extracellular trap and mitophagy-related differentially expressed genes (NETMRDEGs) were used for gene ontology (GO) and pathway (KEGG) enrichment analysis, and the specific results are shown in Table 2.

The results showed that 18 neutrophil extracellular traps and mitophagy-related differentially expressed genes (NETMRDEGs) were mainly enriched in leukocyte aggregation, neutrophil chemotaxis, and neutrophil chemotaxis in polycystic ovary syndrome (PCOS). Leukocyte cell-cell adhesion, leukocyte migration, granulocyte chemotaxis and other biological processes (BP); secretory granule lumen, cytoplasmic vesicle lumen, vesicle lumen, phagocytic vesicle, phagophore assembly site and other cellular components (CC); RAGE receptor binding, Toll-like receptor binding, long-chain fatty acid binding, integrin binding, quaternary ammonium group binding and other molecular functions (MF). Meanwhile, it was also enriched in Yersinia infection, Shigellosis, NOD-like receptor signaling pathway, IL-17 signaling pathway, Lipid and atherosclerosis pathway (KEGG). The results of Gene ontology (GO) and pathway (KEGG) enrichment analysis were visualized by bar graphs (Figure 4A).

https://doi.org/10.2147/IJWH.S560802

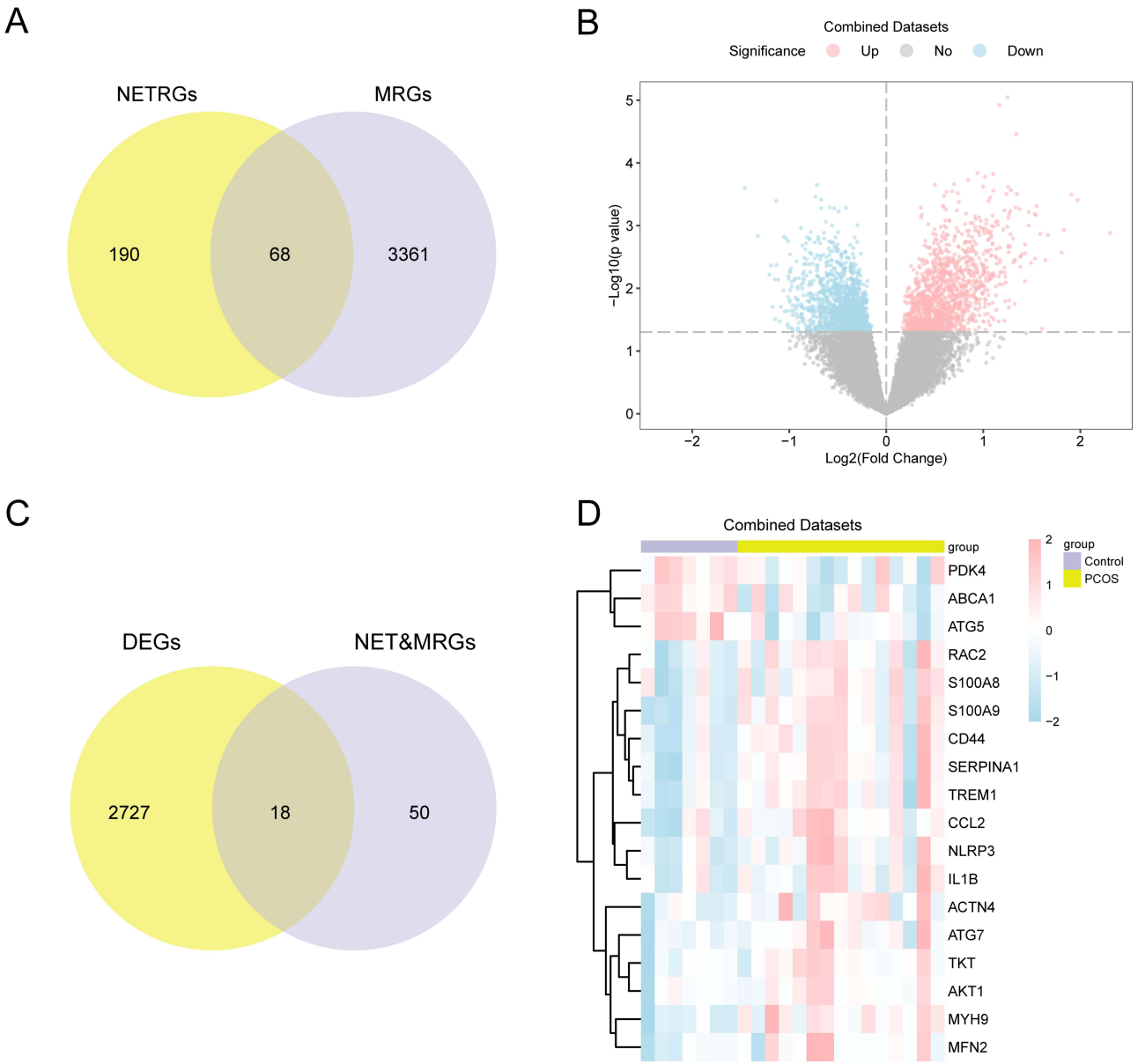

Figure 3 Differential Gene Expression Analysis. (**A**) Venn diagram of neutrophil extracellular trap associated genes (NETRGs) and mitophagy associated genes (MRGs). (**B**) Volcano plot of differentially expressed genes analysis between polycystic ovary syndrome (PCOS) group and Control (Control) group in Combined GEO Datasets. (**C**) Differentially expressed genes (DEGs) and neutrophil extracellular traps and mitophagy-related genes (NETMRGs) Venn plots in the integrated GEO Datasets (Combined Datasets). (**D**) Heat map of neutrophil extracellular traps and mitophagy-related differentially expressed genes (NETMRDEGs) in the integrated GEO Datasets (Combined Datasets). Yellow is polycystic ovary syndrome (PCOS) group, purple is Control (Control) group. Pink represents high expression and blue represents low expression in the heat map.

Abbreviations: PCOS, Polycystic ovary syndrome; NETRGs, Neutrophil Extracellular Traps-Related Genes; MRGs, Mitophagy-Related Genes; DEGs, Differentially Expressed Genes; NETMRGs, Neutrophil Extracellular Traps and Mitophagy-Related Genes; NETMRDEGs, Neutrophil Extracellular Traps and Mitophagy-Related Differentially Expressed Genes.

Meanwhile, the network diagram of biological process (BP), cellular component (CC), molecular function (MF) and biological pathway (KEGG) was drawn according to Gene ontology (GO) and pathway (KEGG) enrichment analysis (Figure 4B–E). The lines show the corresponding molecules and the annotations of the corresponding entries, and the larger the nodes, the more molecules the entries contain.

Gene Set Enrichment Analysis (GSEA)

To determine the effect of expression levels of all genes in Combined GEO Datasets on the pathogenesis of polycystic ovary syndrome (PCOS), Based on the logFC values of all genes in the Combined GEO Datasets between the polycystic

Table 2 Results of GO and KEGG Enrichment Analysis for NETMRDEGs

Ontology	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust
BP	GO:0070486	Leukocyte aggregation	5/18	13/18,800	5.61148E-13	8.04126E-10
BP	GO:0030593	Neutrophil chemotaxis	6/18	106/18,800	4.89263E-10	3.50557E-07
BP	GO:0007159	Leukocyte Cell-cell adhesion	8/18	381/18,800	9.6932E-10	3.67046E-07
BP	GO:0050900	Leukocyte migration	8/18	384/18,800	1.0312E-09	3.67046E-07
BP	GO:0071621	Granulocyte chemotaxis	6/18	128/18,800	1.53683E-09	3.67046E-07
CC	GO:0034774	Secretory granule lumen	5/18	322/19,594	8.35069E-06	0.000278959
CC	GO:0060205	Cytoplasmic vesicle lumen	5/18	325/19,594	8.73491E-06	0.000278959
CC	GO:0031983	Vesicle lumen	5/18	327/19,594	8.99866E-06	0.000278959
CC	GO:0045335	Phagocytic vesicle	3/18	138/19,594	0.000258133	0.006001595
CC	GO:0000407	Phagophore assembly site	2/18	33/19,594	0.000413812	0.006262041
MF	GO:0050786	RAGE receptor binding	2/18	10/18,410	4.04423E-05	0.003377064
MF	GO:0035325	Toll-like receptor binding	2/18	12/18,410	5.92467E-05	0.003377064
MF	GO:0036041	Long-chain fatty acid binding	2/18	15/18,410	9.40925E-05	0.003575514
MF	GO:0005178	Integrin binding	3/18	156/18,410	0.000443554	0.012641285
MF	GO:0050997	Quaternary ammonium group binding	2/18	39/18,410	0.000654858	0.01493076
KEGG	hsa05135	Yersinia infection	5/17	138/8538	5.43033E-06	0.000428502
KEGG	hsa05131	Shigellosis	6/17	250/8538	5.60133E-06	0.000428502
KEGG	hsa04621	NOD-like receptor signaling pathway	5/17	189/8538	2.51275E-05	0.001171978
KEGG	hsa04657	IL-17 signaling pathway	4/17	95/8538	3.06399E-05	0.001171978
KEGG	hsa05417	Lipid and atherosclerosis	5/17	216/8538	4.77634E-05	0.001461561

Abbreviations: GO, Gene Ontology; BP, Biological Process; CC, Cellular Component; MF, Molecular Function; KEGG, Kyoto Encyclopedia of Genes and Genomes; NETMRDEGs, Neutrophil Extracellular Traps and Mitophagy-Related Differentially Expressed Genes.

ovary syndrome (PCOS) group and the Control group, Gene set enrichment analysis (GSEA) was used to investigate the relationship between the expression of all genes in the integrated GEO Datasets (Combined Datasets) and the biological processes, cellular components and molecular functions they played, which were presented by mountain plot (Figure 5A). The detailed results are shown in Table 3. The results showed that all the genes in the Combined Datasets were significantly enriched in Il1 Pathway (Figure 5B), Oxidative Damage Response (Figure 5C), Mapk Pathway (Figure 5D), and the expression of the genes in the combined datasets was significantly enriched. Jak Stat Signaling Pathway (Figure 5E) and other biologically relevant functions and signaling pathways.

Gene Set Variation Analysis (GSVA)

In order to explore the c2. Cp. V2023.2. Hs. Symbols. The GMT gene set in integration of GEO data set (Combined Datasets) of polycystic ovary syndrome (PCOS) and Control (Control) group, the difference between Gene set variation analysis (GSVA) was performed on all genes in the integrated GEO Datasets (Combined Datasets), as detailed in Table 4. Subsequently, the Top20 pathways with p value < 0.05 and in descending logFC absolute value were screened, and the differential expression of the 20 pathways between the polycystic ovary syndrome (PCOS) group and the Control group was analyzed and visualized by heat map (Figure 6A).

Subsequently, the difference was verified based on the Mann–Whitney *U*-test, and the group comparison figure (Figure 6B) was drawn to show the results. The results of gene set variation analysis (GSVA) showed that the pathway autosomal recessive osteopetrosis pathways, medicus reference: BCR-BCAP-CD19-PI3K signaling pathway, induction of autophagy and toll-like receptor signaling pathways by graphene oxide, Biocarta IL-4 pathway, Biocarta IL-10 pathway, Biocarta PCAF pathway, mammary gland development pathway: Involution Stage 4 of 4, Biocarta ERK5 Pathway, Biocarta Srcrptp Pathway, Biocarta BTG2 Pathway, Reactome: The NLRP3 Inflammasome, Reactome: Inflammasomes IL-1 and megakaryocytes in obesity microglia pathogen phagocytosis pathway SA MMP Cytokine Connection, MYD88 Distinct Input-Output Pathway Reactome: Trafficking and Processing of Endosomal TLR, Biocarta B-Lymphocyte Pathway, Propanoate Metabolism were statistically significant in polycystic ovary syndrome (PCOS) group and Control (Control) group (p value < 0.05).

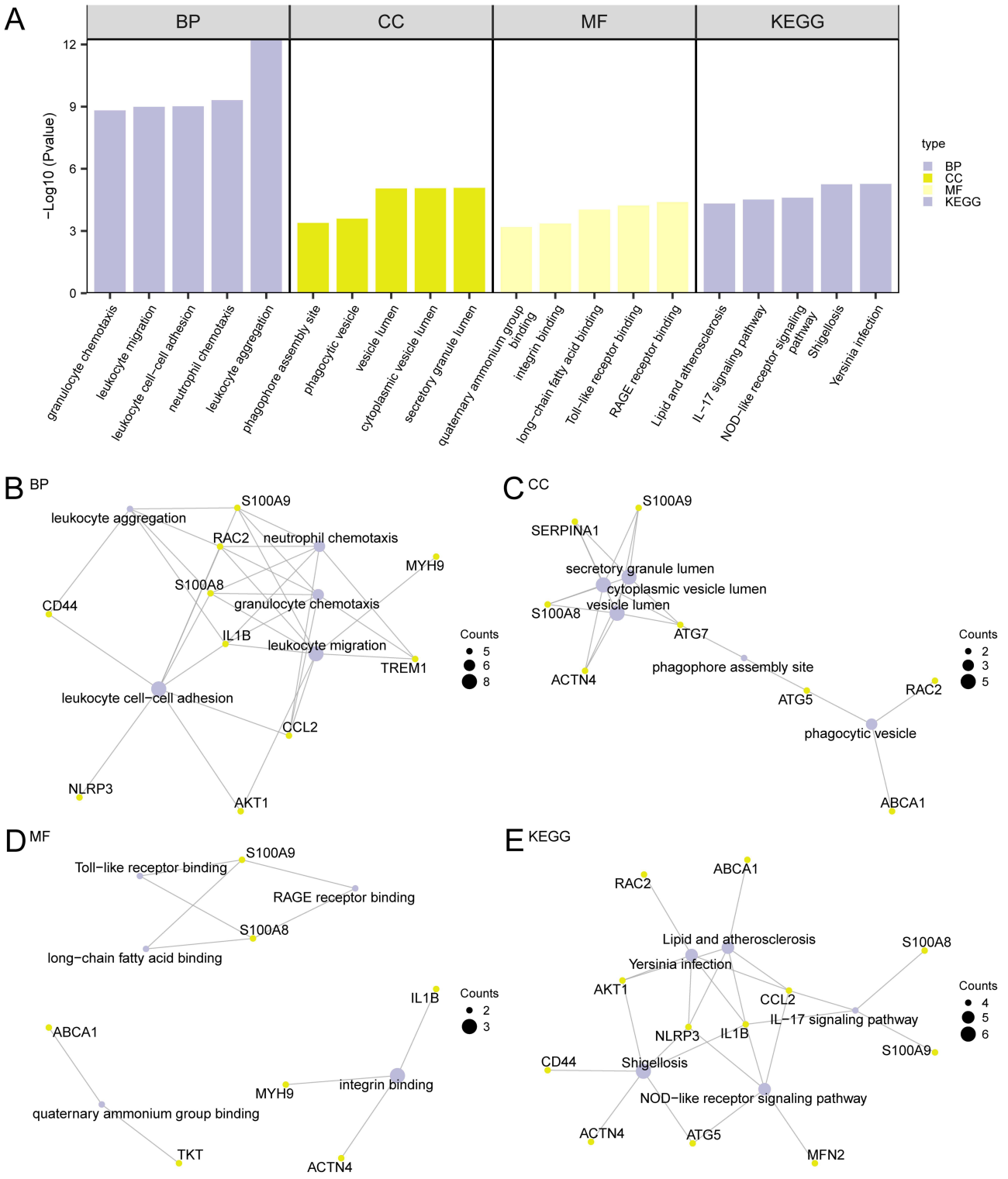

Figure 4 GO and KEGG Enrichment Analysis for NETMRDEGs. (**A**) The bar chart of gene ontology (GO) and pathway (KEGG) enrichment analysis results of neutrophil extracellular trap and mitophagy-related differentially expressed genes (NETMRDEGs) shows: biological process (BP), cellular component (CC), molecular function (MF) and biological pathway (KEGG). GO terms and KEGG terms are shown on the abscissa. (**B**–**E**) Gene ontology (GO) and pathway (KEGG) enrichment analysis results of neutrophil extracellular trap and mitophagy-related differentially expressed genes (NETMRDEGs) network diagram showing BP (**B**), CC (**C**), MF (**D**) and KEGG (**E**). Purple nodes represent items, yellow nodes represent molecules, and the lines represent the relationship between items and molecules. The screening criteria for gene ontology (GO) and pathway (KEGG) enrichment analysis were p value < 0.05 and FDR value (q value) < 0.05, and the p value correction method was Benjamini-Hochberg (BH).
Abbreviations: NETMRDEGs, Neutrophil Extracellular Traps and Mitophagy-Related Differentially Expressed Genes; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; BP, Biological Process; CC, Cellular Component; MF, Molecular Function.

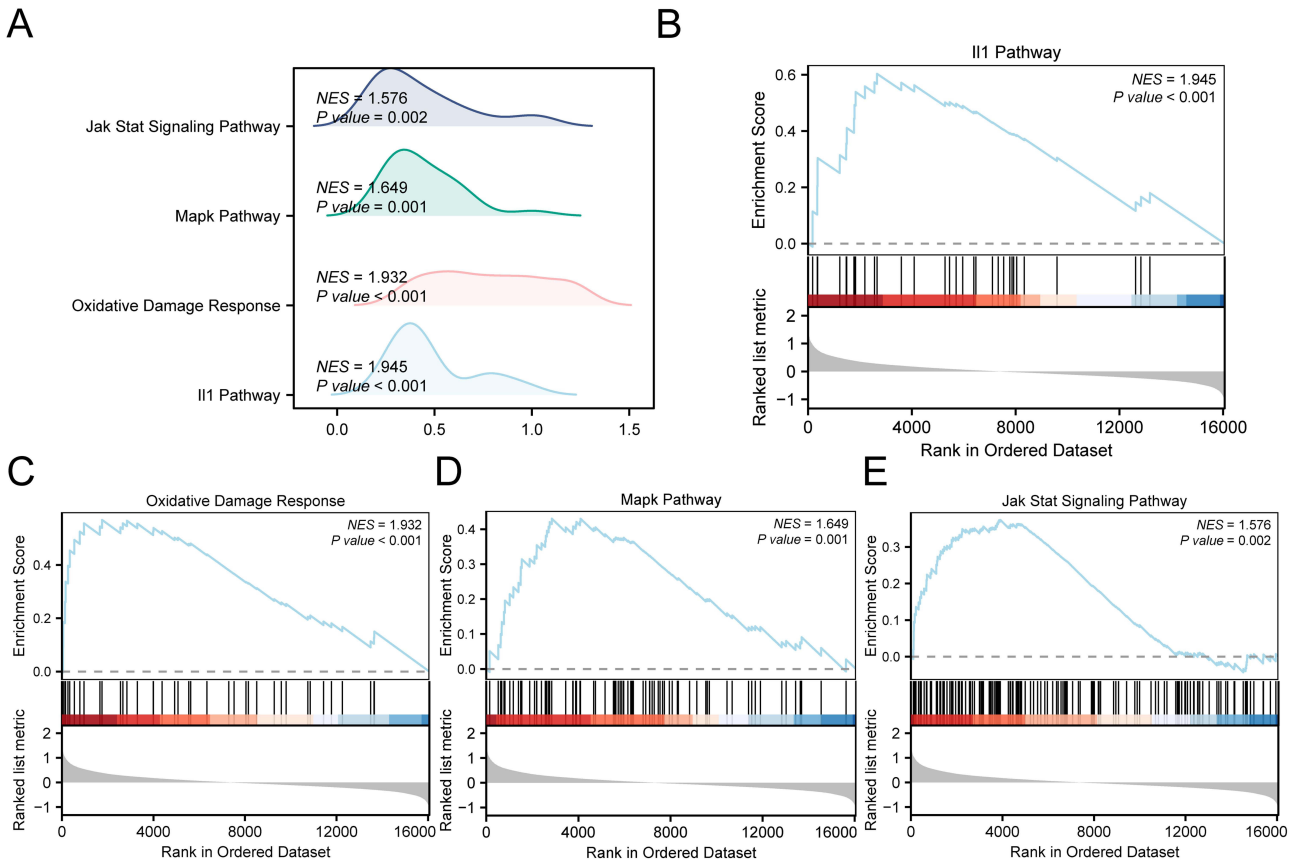

Figure 5 Differential Gene Expression Analysis and GSEA for Combined Datasets. (**A**) Gene set enrichment analysis (GSEA) 4 biological functions mountain map display of the Combined GEO Datasets. (**B**–**E**) Gene set enrichment analysis (GSEA) showed that the integrated GEO Datasets (Combined Datasets) were significantly enriched in Il1 Pathway (**B**), Oxidative Damage Response (**C**), Mapk Pathway (**D**), Jak Stat Signaling Pathway (**E**). GSEA, Gene Set Enrichment Analysis. The screening criteria of gene set enrichment analysis (GSEA) were p value < 0.05 and FDR value (q value) < 0.05.

Construction of Diagnostic Model for Polycystic Ovary Syndrome

Firstly, to determine the diagnostic value of 18 neutrophil extracellular traps and mitophagy-related differentially expressed genes (NETMRDEGs) in polycystic ovary syndrome (PCOS), Univariate logistic regression model was performed based on 18 neutrophil extracellular traps and mitophagy-related differentially expressed genes (NETMRDEGs). The results showed that 10 neutrophil extracellular traps and mitophagy-related differentially expressed genes (NETMRDEGs) were statistically significant in the Logistic regression model (p value < 0.05), see Table S4.

Then, based on the 18 neutrophil extracellular traps and mitophagy-related differentially expressed genes (NETMRDEGs), the Least Absolute Shrinkage and Selection (LASSO) was performed Operator regression analysis was used to construct the LASSO regression model, namely the diagnosis model of polycystic ovary syndrome (PCOS). LASSO variable trajectory diagram (Figure 7A) and LASSO regression model diagram (Figure 7B) were drawn for visualization.

Table 3 Results of GSEA for Combined Datasets

ID	Set Size	Enrichment Score	NES	pvalue	p.adjust	qvalue
PID_IL1_PATHWAY	32	0.603341538	1.944900131	0.000200365	0.003089378	0.002480835
WP_OXIDATIVE_DAMAGE_RESPONSE	39	0.569128511	1.931574467	0.000203105	0.003098978	0.002488545
BIOCARTA_MAPK_PATHWAY	80	0.430397455	1.649118686	0.001111929	0.012722749	0.010216633
KEGG_JAK_STAT_SIGNALING_PATHWAY	146	0.374087723	1.576387393	0.001839978	0.018603385	0.014938908

Abbreviation: GSEA, Gene Set Enrichment Analysis.

https://doi.org/10.2147/IJWH.S560802

Table 4 Results of GSVA for Combined Datasets

Pathway	logFC	AveExpr	t	P.Value	adj.P.Val	B
Biocarta Blymphocyte Pathway	0.906675116	0.010693081	3.633942733	0.000661688	0.039232645	−0.423496922
WP III and Megakaryocytes in Obesity	0.897801829	0.021668962	4.413432815	5.49E-05	0.019596743	1.781852328
SA MMP Cytokine Connection	0.864453259	0.021677058	3.879544718	0.00030874	0.031778971	0.249199103
WP Microglia Pathogen Phagocytosis Pathway	0.854654369	0.023802762	4.109507399	0.000148325	0.026044172	0.898588402
WP Mammary Gland Development Pathway Involution Stage 4 of 4	0.844339541	−0.005205776	4.47233486	4.51E-05	0.019596743	1.955945692
WP Autosomal Recessive Osteopetrosis Pathways	0.841257139	−0.012430396	4.428403598	5.22E-05	0.019596743	1.826017089
Reactome Trafficking and Processing of Endosomal TLR	0.824889149	0.08081693	3.647527768	0.000634747	0.038926292	−0.386897929
Biocarta IL10 Pathway	0.817229952	−0.031082593	3.721441681	0.000505634	0.034643116	−0.186483703
WP MYD88 Distinct Input Output Pathway	0.80840221	−0.00909147	3.950638096	0.000246599	0.030670693	0.44805651
Kegg Medicus Reference BCR BCAP CD19 PI3K Signaling Pathway	0.805843009	−0.020482691	4.27686431	8.61E-05	0.019596743	1.381711743
Reactome Inflammasomes	0.802263555	0.038102258	3.975445328	0.000227906	0.029628076	0.517856342
WP Induction of Autophagy and Toll like Receptor Signaling Pathways by Graphene Oxide	0.799771648	−0.02151536	4.066845359	0.000170158	0.026732686	0.776798702
Biocarta IL4 Pathway	0.79731169	0.005641153	4.227795115	0.00010106	0.019596743	1.239202774
Biocarta ERK5 Pathway	0.794163167	−0.005268985	4.26504848	8.95E-05	0.019596743	1.347332763
Reactome the NLRP3 Inflammasome	0.785553252	0.033706374	3.992479112	0.000215872	0.029628076	0.56590462
Kegg Propanoate Metabolism	−0.781225585	0.002578693	−4.44135404	5.00E-05	0.019596743	1.864267952
Biocarta PCAF Pathway	0.777968174	−0.032656286	4.089857049	0.000158021	0.026205183	0.842419905
Biocarta TCRA Pathway	0.773146011	−0.010241424	2.906240285	0.005452984	0.074960708	−2.261915863
Biocarta SRCRPTP Pathway	0.772799634	0.027651682	4.3731701	6.27E-05	0.019596743	1.663363512
Biocarta BTG2 Pathway	0.769504736	−0.013237518	3.905819075	0.000284191	0.031312533	0.322485951

https://doi.org/10.2147/IJWH.S560802

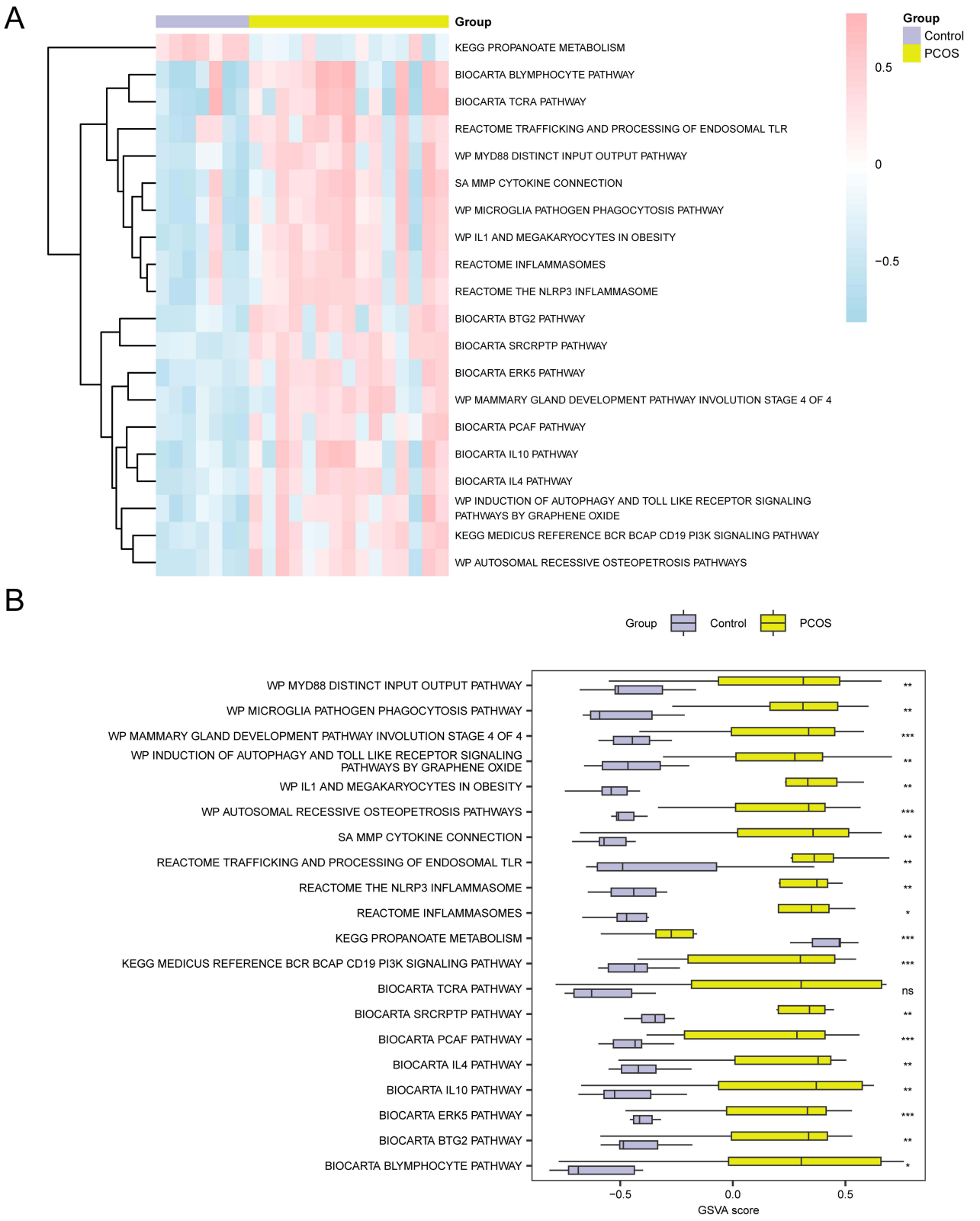

Figure 6 GSVA Analysis. (**A** and **B**) Heat map (**A**) and group comparison map (**B**) of gene set variation analysis (GSVA) results between polycystic ovary syndrome (PCOS) and Control groups in Combined GEO Datasets. Polycystic ovary syndrome, Sepsis-induced cardiomyopathy; GSVA, Gene Set Variation Analysis. ns stands for p value ≥ 0.05, not statistically significant; * represents p value < 0.05, statistically significant; ** represents p value < 0.01, highly statistically significant; *** represents p value < 0.001 and highly statistically significant. Yellow represents the polycystic ovary syndrome (PCOS) group and purple represents the Control (Control) group. The screening criterion for gene set variation analysis (GSVA) was a P-value < 0.05. Blue represents low enrichment and pink represents high enrichment in the heat map.

https://doi.org/10.2147/IJWH.S560802

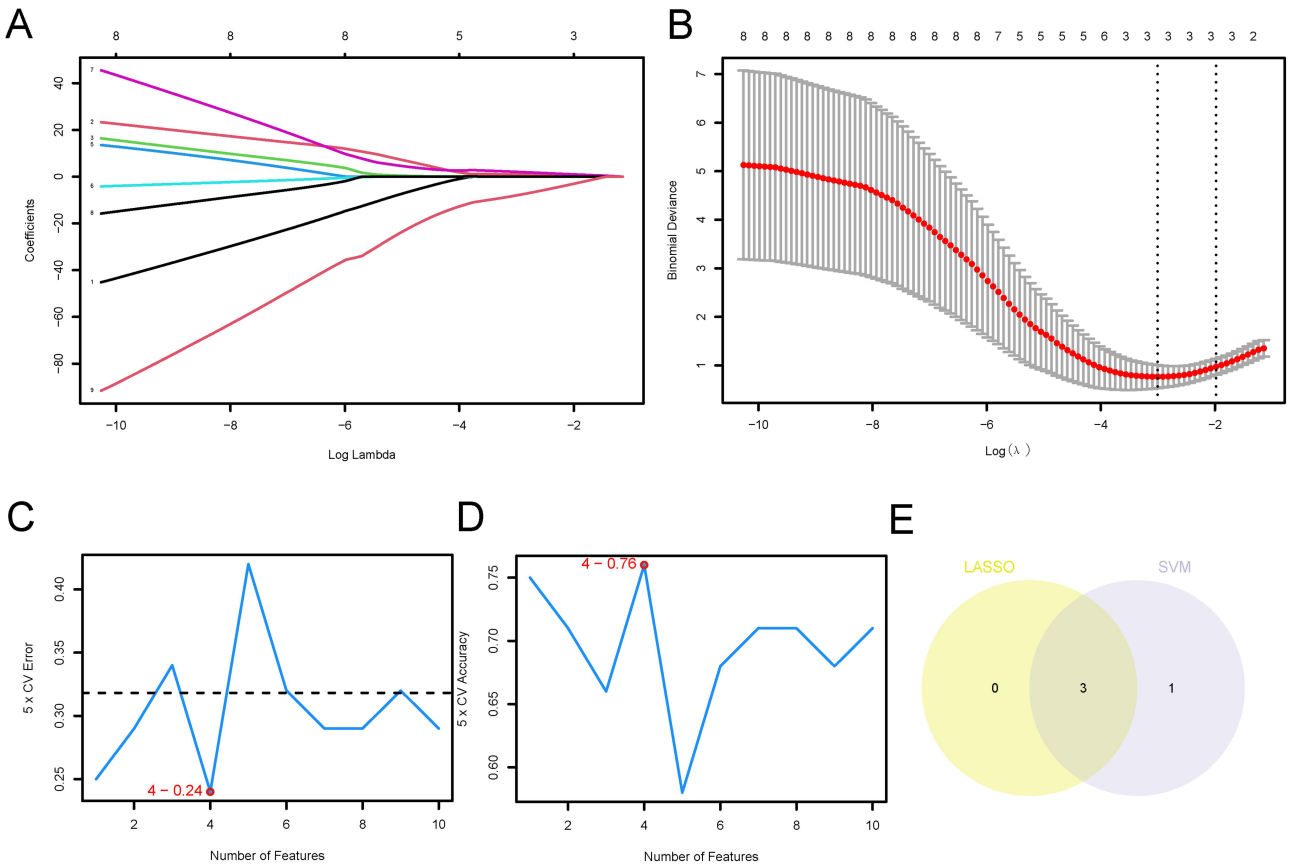

Figure 7 Construction of diagnostic model for polycystic ovary syndrome. (**A**) LASSO regression variable trajectories of neutrophil extracellular traps and mitophagy-related differentially expressed genes (NETMRDEGs) in the Combined GEO Datasets. (**B**) Regression model plot of the LASSO diagnostic model. (**C**) The number of genes with the lowest error rate obtained by the SVM algorithm. (**D**) The number of genes with the highest accuracy obtained by the SVM algorithm. (**E**) Venn diagram of intersection between LASSO algorithm and SVM algorithm.

Abbreviations: LASSO, Least Absolute Shrinkage and Selection Operator; SVM, Support Vector Machine; NETMRDEGs, Neutrophil Extracellular Traps and Mitophagy-Related Differentially Expressed Genes.

The results showed that three neutrophil extracellular trap and mitophagy-related differentially expressed Genes (NETMRDEGs), namely Model Genes, included in the LASSO regression model were S100A9, MYH9 and ATG5.

Finally, the SVM model was constructed based on 18 neutrophil extracellular trap and mitophagy-related differentially expressed genes (NETMRDEGs) and SVM (Support Vector Machine) algorithm, and the number of genes with the lowest error rate (Figure 7C) and the highest accuracy (Figure 7D) was obtained. The results showed that the SVM model had the highest accuracy when the number of genes was 4, and the four neutrophil extracellular trap and mitophagy-related differentially expressed genes (NETMRDEGs) were MYH9, S100A9, ATG5 and SERPINA1, respectively.

In order to obtain the Key Genes, the intersection of neutrophil extracellular traps and mitophagy related differentially expressed genes (NETMRDEGs) in LASSO regression model and neutrophil extracellular traps and mitophagy related differentially expressed genes (NETMRDEGs) in SVM model was taken. A total of 3 neutrophil extracellular traps and mitophagy-related differentially expressed Genes (NETMRDEGs) were obtained as Key Genes for subsequent study and Venn diagram (Figure 7E) was drawn. The 3 Key Genes were S100A9, MYH9 and ATG5.

Validation of Diagnostic Model for Polycystic Ovary Syndrome

To further validate the diagnostic model of polycystic ovary syndrome (PCOS), a Nomogram based on Key Genes was used to show the interrelationships of Key Genes in the Combined GEO Datasets (Figure 8A). The results showed that the expression level of the Key Genes MYH9 had significantly higher utility than other variables in the diagnostic model of polycystic ovary syndrome (PCOS).

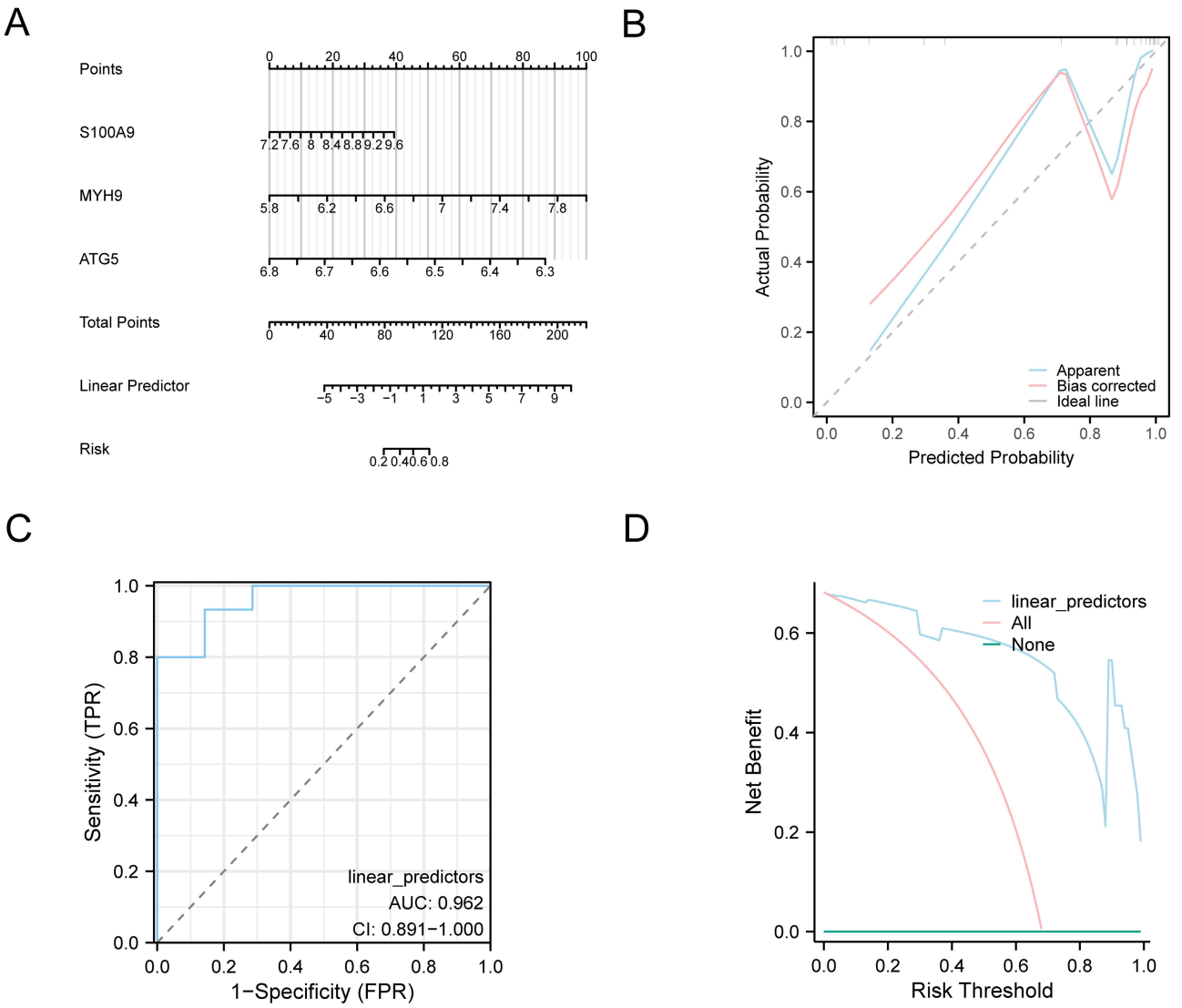

Figure 8 Diagnostic and Validation Analysis of PCOS. (**A**) Nomograms of Key Genes in Combined GEO Datasets in polycystic ovary syndrome (PCOS) diagnostic models. (**B** and **C**) Calibration Curve plot (**B**) and decision curve analysis (DCA) plot (**C**) of Key Genes in integrated GEO Datasets for polycystic ovary syndrome (PCOS) diagnostic model. (**D**) ROC analysis of Linear Predictors of Logistic regression models in GEO Datasets (Combined Datasets). The ordinate of decision curve analysis (DCA) plot is the net benefit, and the abscissa is the Probability Threshold or Threshold Probability; AUC > 0.9 had high accuracy.
Abbreviations: DCA, Decision Curve Analysis; ROC, Receiver Operating Characteristic; AUC, Area Under the Curve.

Then, in order to judge the accuracy and discrimination of the diagnostic model for polycystic ovary syndrome (PCOS), a Calibration Curve was drawn by Calibration analysis. The predictive effect of the model on the actual results was evaluated according to the fitting of the actual probability and the predicted probability of the model in different situations drawn in the figure (Figure 8B). The Calibration Curve plot of the diagnostic model for polycystic ovary syndrome (PCOS) shows that the calibration line shown by the dotted line is slightly deviated from the diagonal line of the ideal model, but close to the fit.

To evaluate and present the clinical utility of polycystic ovary syndrome (PCOS) diagnostic models by decision curve analysis (DCA) based on the Key Genes in the Combined GEO Datasets (Figure 8C). The results show that the line of the model is stable higher than that of All positive and All negative in a certain range, and the net benefit of the model is more, and the effect of the model is better.

We also plotted the Receiver Operating characteristics of the Linear Predictors of the Logistic regression model across different groups (PCOS/Control) in the Combined GEO Datasets The receiver operating characteristic (ROC) Curve was plotted and the results were presented (Figure 8D). The figure shows that the Logistic regression model of the Combined GEO Datasets has a better diagnostic effect.

https://doi.org/10.2147/IJWH.S560802

Gene Set Enrichment Analysis (GSEA)

The disease group (PCOS) samples of Combined GEO Datasets were divided into HighRisk group and LowRisk group according to the LASSO RiskScore calculated by the risk coefficient of LASSO regression analysis. To analyze the differences in gene expression values between the HighRisk and LowRisk groups in the Combined GEO Datasets, The R package limma was used to perform differential analysis on the Combined GEO Datasets to obtain the differentially expressed genes between the two groups.

To determine the effect of expression levels of all genes in the integrated GEO Datasets (Combined Datasets) on polycystic ovary syndrome (PCOS) pathogenesis, Based on the logFC values of all genes in the Combined GEO Datasets between the HighRisk and LowRisk groups, Gene set enrichment analysis (GSEA) was used to investigate the relationship between the expression of all genes in the Combined GEO Datasets and the involved biological processes, cellular components and molecular functions, which were presented by mountain plot (Figure 9A). The results showed that all genes in the integrated GEO Datasets (Combined Datasets) were significantly enriched in Kegg Jak Stat Signaling Pathway (Figure 9B), Pid Pi3kci Pathway (Figure 9C), and KEGG. Ryan Mantle Cell Lymphoma Notch Direct Up (Figure 9D), Stambolsky Targets Of Mutated Tp53 Dn (Figure 9E) and other biologically relevant functions and signaling pathways.

Gene Set Variation Analysis (GSVA)

To explore the c2. Cp. V2023.2. Hs. Symbols. The GMT gene set in integration of GEO data set (Combined Datasets) risky (HighRisk) group and low risk (LowRisk) group, the difference between Gene set variation analysis (GSVA) was performed on all genes in the Combined GEO Datasets.

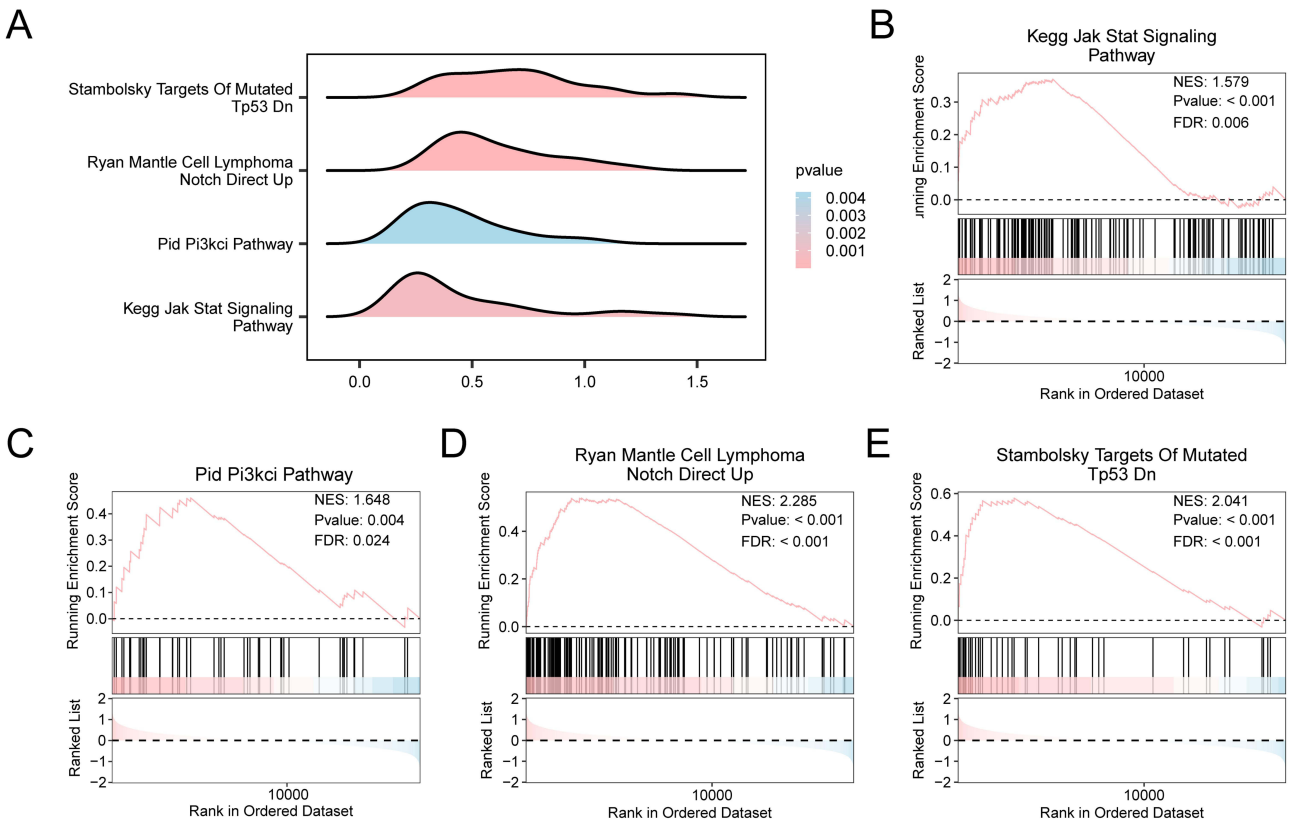

Figure 9 Differential Gene Expression Analysis and GSEA for Combined Datasets. (**A**) 4 biological functions mountain map of gene set enrichment analysis (GSEA) of Combined GEO Datasets. (**B**–**E**) Gene set enrichment analysis (GSEA) showed that the integrated GEO Datasets (Combined Datasets) were significantly enriched in Kegg Jak Stat Signaling Pathway (**B**), Pid Pi3kci Pathway (**C**), Ryan Mantle Cell Lymphoma Notch Direct Up (**D**), Stambolsky Targets Of Mutated Tp53 Dn (**E**). The screening criteria of gene set enrichment analysis (GSEA) were p value < 0.05 and FDR value (q value) < 0.05.
Abbreviation: GSEA, Gene Set Enrichment Analysis.

https://doi.org/10.2147/IJWH.S560802

Subsequently, the Top20 pathways with p value < 0.05 and in descending logFC absolute value were screened, and the differential expression of 20 pathways between the HighRisk group and the LowRisk group was analyzed and visualized by heat map (Figure 10A).

Then, the differences were verified based on the Mann–Whitney *U*-test, and the group comparison chart (Figure 10B) was drawn to show the results. The results OF gene set variation analysis (GSVA) showed that Steroid Biosynthesis, Biosynthesis of Unsaturated Fatty Acids, Mevalonate Arm of Cholesterol Biosynthesis Pathway, Medicus Reference Cholesterol Biosynthesis, TCA Cycle and Deficiency of Pyruvate Dehydrogenase Complex (PDHC), Omega-9 Fatty Acid Synthesis, Terpenoid Backbone Biosynthesis, Medicus Reference Mevalonate Pathway, Reactome Cholesterol Biosynthesis, Enterocyte Cholesterol Metabolism, Cholesterol Synthesis Disorders, Cholesterol Biosynthesis Pathway, Medicus Variant Scrapie Conformation PrPsc to Prnp PI3K NOX2 Signaling Pathway, Biocarta Lym Pathway, Platelet Mediated Interactions with Vascular and Circulating Cells, Biocarta T-Helper Pathway, Biocarta T-Cytotoxic Pathway, Biocarta Ctl Pathway, Biocarta Tcra Pathway, The Biocarta Monocyte Pathway was statistically significant in the HighRisk group and the LowRisk group ($p < 0.05$).

Differential Expression Analysis, Correlation Analysis and ROC Curve Analysis

To explore the expression differences of three Key Genes between polycystic ovary syndrome (PCOS) group and Control group in the Combined GEO Datasets, we show by drawing group comparison diagram (Figure 11A), and the results show that: The results showed that the expression levels of ATG5 and S100A9 were highly statistically significant ($p < 0.01$); The expression of MYH9 was highly statistically significant ($p < 0.001$).

The R package pROC was used to draw ROC curves based on the expression levels of Key Genes in the Combined GEO Datasets. The ROC curve (Figure 11B–D) showed that the expression value of ATG5 had a certain accuracy in the diagnosis of Control group ($0.7 < AUC < 0.9$) and polycystic ovary syndrome (PCOS) group ($0.7 < AUC < 0.9$); The expression values of MYH9 and S100A9 had high diagnostic accuracy ($AUC > 0.9$) for the Control group and polycystic ovary syndrome (PCOS) group.

Correlation analysis and correlation heatmap were performed based on the complete expression matrix of the three Key Genes in the Combined GEO Datasets (Figure 12A). The results showed that there was a certain positive correlation between the Key Genes MYH9 and S100A9. The Key Genes S100A9 and ATG5 were negatively correlated.

The locations of three Key Genes on human chromosome were annotated and visualized by circle diagram (Figure 12B). The figure showed that the Key Genes S100A9 were located on chromosome 1, ATG5 on chromosome 6, and MYH9 on chromosome 22.

Immune Infiltration Analysis

The expression matrices of Combined GEO Datasets were applied to calculate the immune infiltration abundance of 28 immune cells by the ssGSEA algorithm. Firstly, the expression differences of infiltrating abundance of immune cells in different groups were shown by group comparison plots. The group comparison diagram (Figure 13A) showed that all the 14 immune cells were statistically significant (p value < 0.05), namely: Activated B cell, Activated CD8 T cell, Activated dendritic cell, Central memory CD4 T cell, Effector memory CD8 T cell, Eosinophil, Immature B cell, Macrophage, Mast cell, MDSC, Natural killer cell, Natural killer T cell, Regulatory T cell, Type 1 T helper cell. Then, the correlation heatmap was used to show the correlation results of the 14 immune cell infiltration abundance in the immune infiltration analysis in the Combined GEO Datasets (Figure 13B). The results showed that most of the immune cells showed strong positive correlations, and the immune cell Activated B cell and Effector memory CD8 T cell had the most significant positive correlation (r value = 0.968, p value < 0.05).

Finally, the correlation between Key Genes and the abundance of immune cell infiltration was shown by correlation bubble plot (Figure 13C). The results of correlation bubble plot showed that most of the immune cells showed strong positive correlation, and S100A9 and immune cell Activated B cell had the strongest significant positive correlation (r value = 0.948, p value < 0.05).

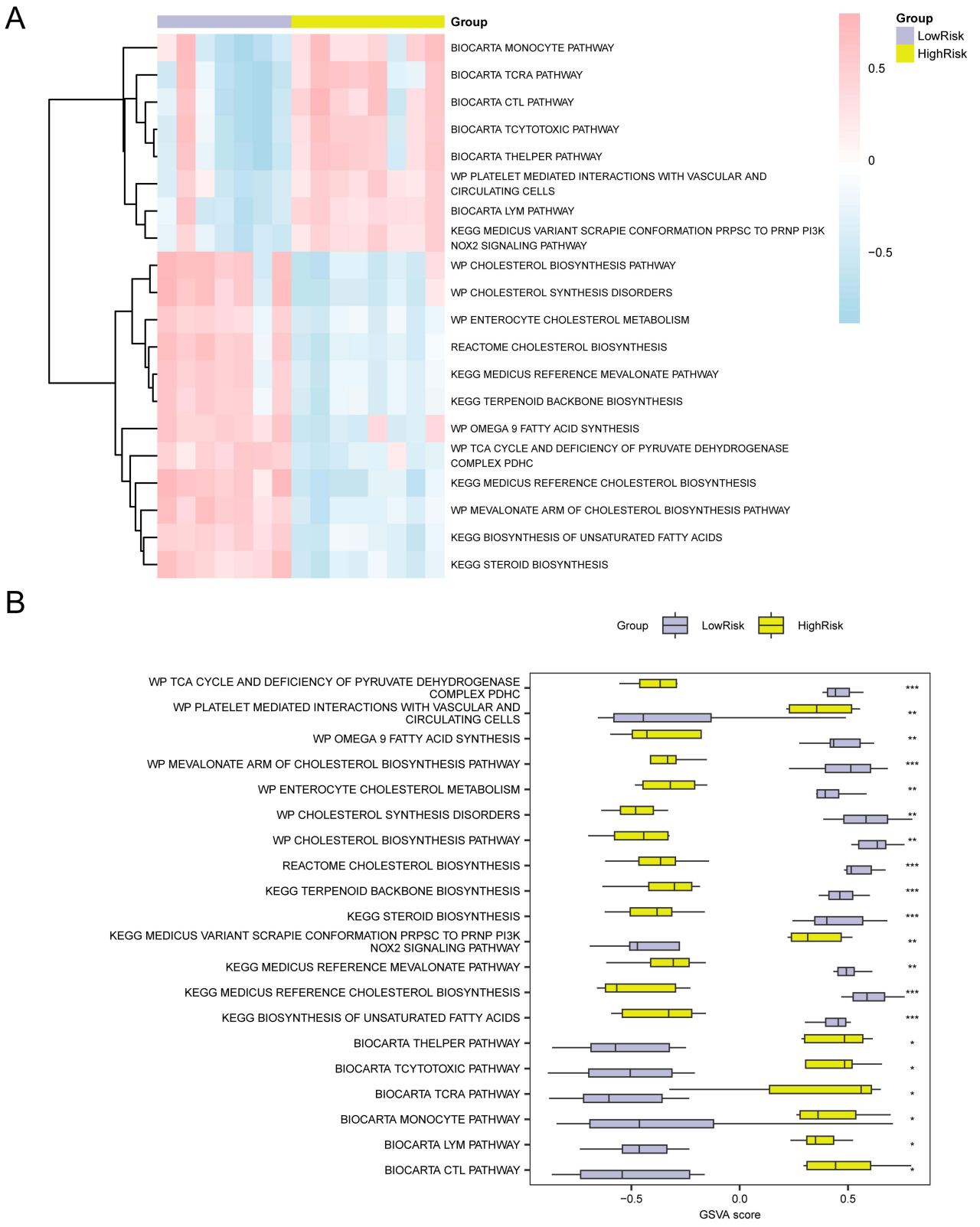

Figure 10 GSVA Analysis. (**A** and **B**) Heat map (**A**) and group comparison map (**B**) of gene set variation analysis (GSVA) results between HighRisk and LowRisk groups of Combined GEO Datasets. ns stands for p value ≥ 0.05, not statistically significant; * represents p value < 0.05, statistically significant; ** represents p value < 0.01, highly statistically significant; *** represents p value < 0.001 and highly statistically significant. Yellow represents the HighRisk group and purple represents the LowRisk group. The screening criteria for gene set variation analysis (GSVA) was p value < 0.05, and the p value correction method was Benjamini-Hochberg (BH). Blue represents low enrichment and pink represents high enrichment in the heat map.

Abbreviations: PCOS, Polycystic ovary syndrome; GSVA, Gene Set Variation Analysis.

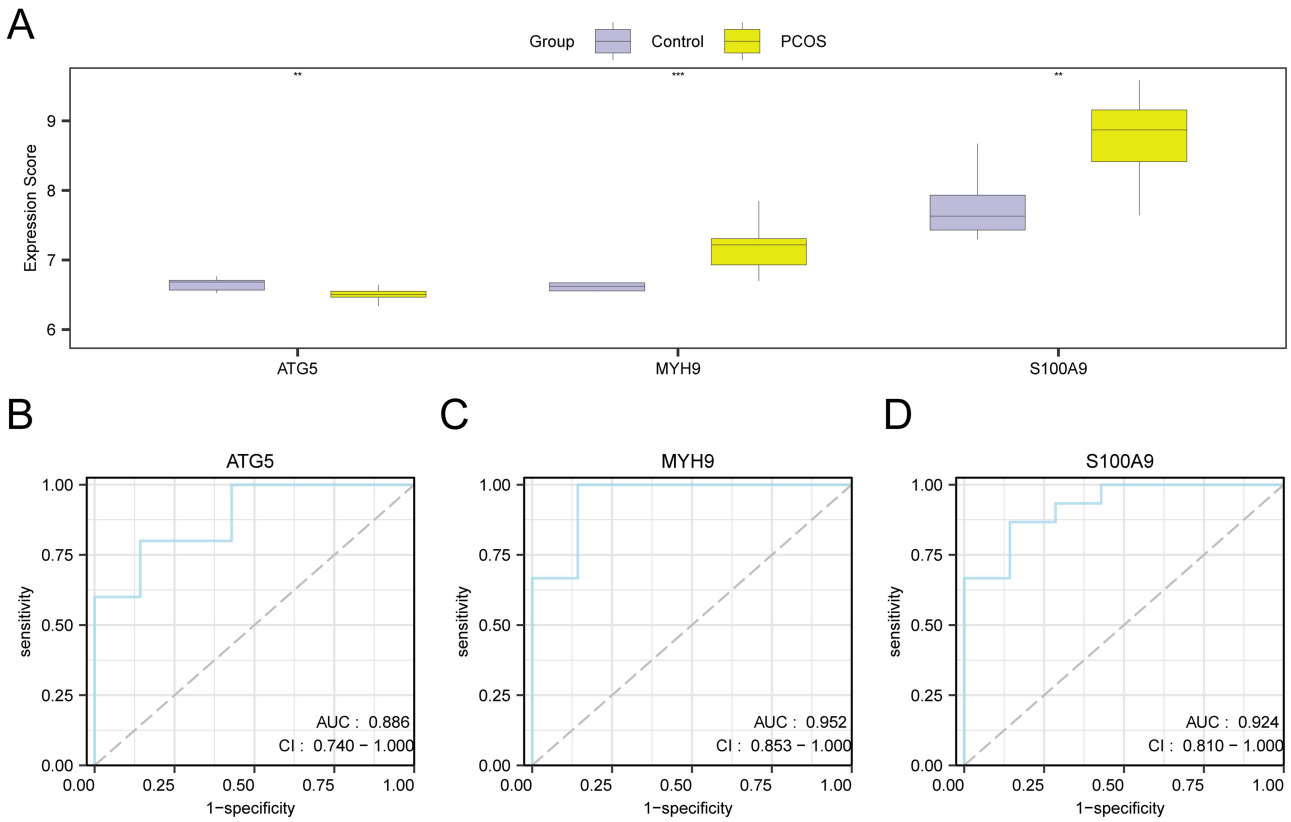

Figure 11 Differential Expression, Correlation Analysis and ROC Curve. (**A**) Group comparison diagram of Key Genes in the polycystic ovary syndrome (PCOS) group and Control group in the Combined GEO Datasets. (**B**–**D**) ROC curves of Key Genes ATG5 (**B**), MYH9 (**C**), S100A9 (**D**) in the Combined GEO Datasets. *** represents p value < 0.001, highly statistically significant. ** represents a p value < 0.01 and is highly statistically significant. When AUC > 0.5, it indicates that the expression of the molecule is a trend to promote the occurrence of the event, and the closer the AUC is to 1, the better the diagnostic effect. The AUC had some accuracy in the range of 0.7 to 0.9. AUC > 0.9 had high accuracy. In the group comparison plot, the polycystic ovary syndrome (PCOS) group is yellow, and the Control group is purple.
Abbreviation: PCOS, Polycystic ovary syndrome.

Protein-Protein Interaction (PPI) Network

Protein-protein interaction analysis of three Key Genes (ATG5, S100A9, MYH9) was performed using STRING database with minimum required interaction score greater than 0.150 (PPI network: low confidence (0.150)) was used as the standard, and the Protein-Protein Interaction network (PPI Network) diagram (Figure 14A) was constructed.

Finally, we predicted the similar Genes of three Key Genes (ATG5, S100A9, MYH9) through GeneMANIA website and drew the interaction network to observe the physical interaction, shared protein domains, gene interaction and other information between them (Figure 14B).

Construction of Regulatory Network

Firstly, the transcription factors (TFS) combined with Key Genes were obtained from ChIPBase database, and the mRNA-TF Regulatory Network was constructed and visualized by Cytoscape software (Figure 15A). Among them, there are 3 Key Genes and 103 transcription factors (TFS), the specific information is shown in Table S5.

Then, the miRNAs related to the Key Genes were obtained from TarBase database, and the mRNA-miRNA Regulatory Network was constructed and visualized by Cytoscape software (Figure 15B). Among them, 2 Key Genes and 120 miRNAs were included, and the specific information is shown in Table S6.

Then, the RNA-binding proteins (RBP) associated with the Key Genes were predicted by StarBase database, and the mRNA-RBP Regulatory Network was constructed and visualized by Cytoscape software (Figure 16A). Among them, there are 2 Key Genes and 108 RNA binding proteins (RBP), the specific information is shown in Table S7.

https://doi.org/10.2147/IJWH.S560802

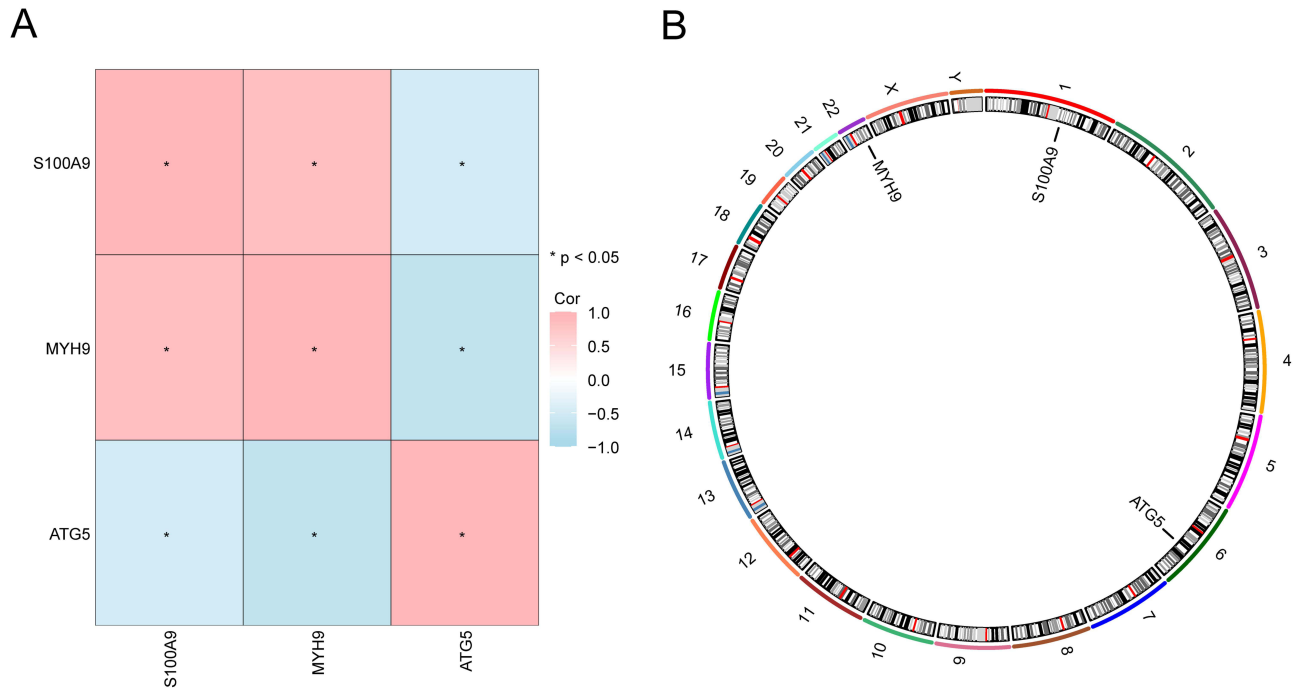

Figure 12 Correlation analysis and chromosome localization analysis of Key Genes. (**A**) Correlation analysis between Key Genes. (**B**) Chromosomal mapping of Key Genes in human body. The absolute value of correlation coefficient (r value) below 0.3 was weak or no correlation, 0.3 to 0.5 was weak correlation, 0.5 to 0.8 was moderate correlation, and above 0.8 was strong correlation.

Finally, the potential drugs or molecular compounds related to the Key Genes were identified by using the CTD database. The mRNA-Drug Regulatory Network (mRNA-Drug Regulatory Network) was constructed and visualized by Cytoscape (Figure 16B). Among them, 3 Key Genes and 20 drugs or molecular compounds were included, as detailed in Table S8.

Discussion

Roles of NETMRDEGs in PCOS Pathogenesis

This study identified 18 NETMRDEGs (including *S100A9, MYH9*, and *ATG5*), which are significantly enriched in inflammatory pathways. The elevated NETs markers (S100A9 and MPO-DNA) observed in PCOS patients corroborate earlier reports linking NETosis to chronic inflammatory conditions[5,6] and suggest that these genes orchestrate PCOS pathogenesis through coordinated inflammation and mitochondrial dysfunction. Specifically, S100A9 encodes a pro-inflammatory alarmin linked to neutrophil activation, while ATG5 is critical for mitophagy, highlighting its dual role in inflammation and metabolic regulation. Notably, S100A9 and MYH9 are implicated in autoimmune diseases, suggesting shared inflammatory mechanisms. ATG5 deficiency disrupts mitochondrial clearance and promotes oxidative stress and follicular atresia, which is consistent with elevated apoptosis observed in PCOS granulosa cells. The key genes (S100A9, MYH9, and ATG5) identified in this study hold significant biological implications in the occurrence and progression of polycystic ovary syndrome (PCOS). Specifically, S100A9 is closely associated with neutrophil activation and the formation of neutrophil extracellular traps (NETs), which can amplify local ovarian inflammation and oxidative stress. MYH9 is involved in cytoskeletal remodeling, and its dysregulation may impair granulosa cell function and follicular development. ATG5 is a key regulator of mitophagy and the maintenance of mitochondrial homeostasis. Collectively, these three genes link amplified inflammation, mitochondrial dysfunction, and ovarian functional impairment, providing a molecular basis for the immunometabolic mechanisms and potential diagnostic and therapeutic targets of PCOS. Future studies should validate therapeutic strategies targeting these genes, such as CRISPR-based S100A9 knockout to confirm its role in androgen-driven NETosis or exploring mitophagy enhancers to restore mitochondrial homeostasis.

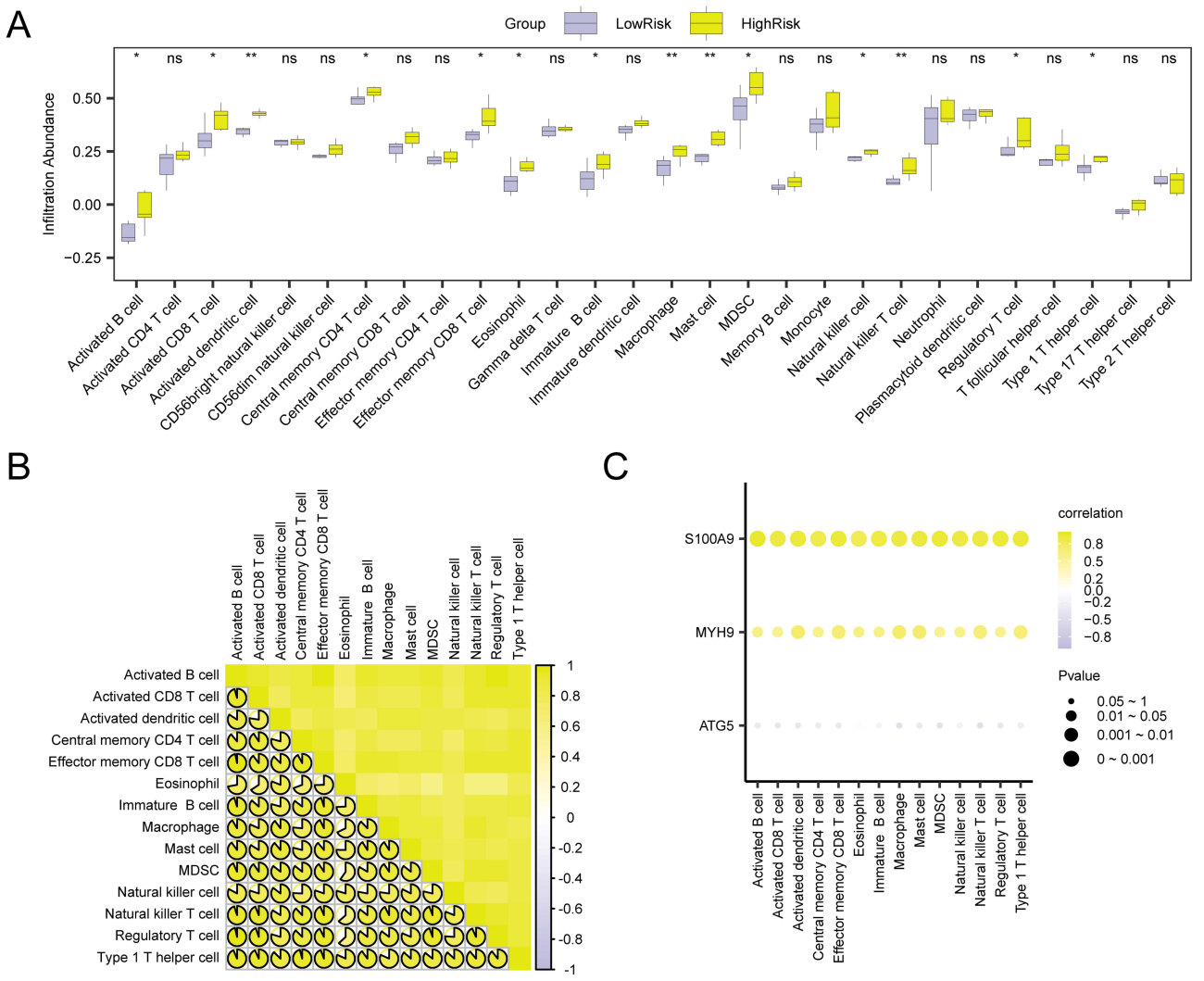

Figure 13 Immune Infiltration Analysis by ssGSEA Algorithm. (**A**) Group comparison diagram of immune cells in the LowRisk group and HighRisk group of the Combined GEO Datasets. (**B**) Correlation heatmap of immune cell infiltration abundance in the integrated GEO Datasets (Combined Datasets). (**C**) Bubble plot of correlation between Key Genes and immune cell infiltration abundance in the Combined GEO Datasets. ns stands for p value ≥ 0.05, no statistical significance; * represents p value < 0.05, indicating statistical significance; ** represents p value < 0.01, highly statistically significant; The absolute value of correlation coefficient (r value) below 0.3 was weak or no correlation, 0.3 to 0.5 was weak correlation, 0.5 to 0.8 was moderate correlation, and above 0.8 was strong correlation. Purple is the LowRisk group, and yellow is the HighRisk group. Yellow is the positive correlation, purple is the negative correlation. The depth of the color represents the strength of the correlation.
Abbreviations: ssGSEA, single-sample Gene-Set Enrichment Analysis; PCOS, Polycystic ovary syndrome.

Immune and Metabolic Crosstalk

According to enrichment analysis, NETMRDEGs were predominantly associated with leukocyte aggregation, neutrophil chemotaxis, and IL-17 signaling, which are critical to chronic inflammation and insulin resistance in PCOS. IL-17 drives ovarian inflammation and metabolic dysfunction, while neutrophil chemotaxis amplifies tissue damage via the release of NETs, consistent with elevated oxidative stress observed in PCOS granulosa cells. Therapeutic strategies targeting these pathways (such as IL-17 inhibitors or CXCR2 antagonists) may mitigate inflammation, particularly in inflammatory subtypes (Cluster 2) characterized by Th17 polarization and elevated gamma delta T cells. Notably, the IL-17/neutrophil axis is central to psoriasis and inflammatory bowel disease, suggesting opportunities to repurpose immunomodulatory therapies for PCOS. Cluster-based stratification further highlights the need for subtype-specific interventions, which aligns with findings that immune microenvironment heterogeneity influences PCOS progression and treatment response.

https://doi.org/10.2147/IJWH.S560802

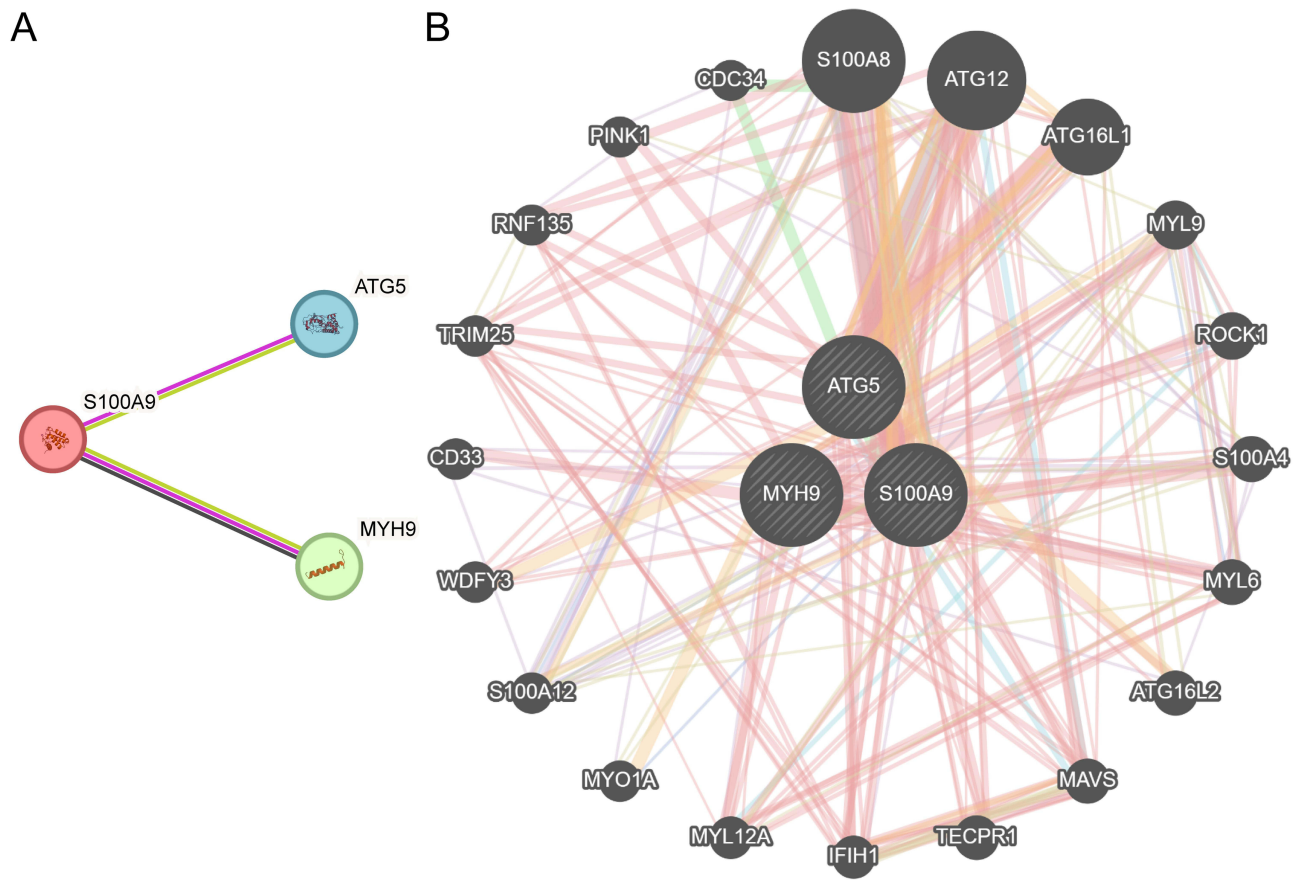

Figure 14 Protein-protein Interaction (PPI) Network. (**A**) PPI Network of Key Genes. (**B**) Key Genes predict the interaction network of genes with similar functions. Circular nodes represent genes, and the size is determined by the attributes and characteristics of the genes. Lines represent relationships, interactions, or functional connections between genes, and line thicknesses represent strong associations or important interactions. PPI Network: Protein-Protein Interaction Network.

Discussion on the Diagnostic Model: Clinical Translation and Optimization

The LASSO regression model incorporating S100A9, MYH9, and ATG5 demonstrated high diagnostic accuracy (AUC > 0.9), outperforming traditional biomarkers (such as testosterone), offering a non-invasive tool for early PCOS detection, especially in atypical cases such as non-obese adolescents. External validation in multi-ethnic cohorts (eg, European and African populations) is critical for addressing genetic bias in current Asian-dominant datasets to enhance clinical utility, as highlighted by Tan et al (2024). Multi-modal integration of gene expression with metabolic markers or imaging features could further improve specificity, aligning with findings that combined biomarkers enhance PCOS subtyping. Additionally, cross-disease application of this model in obesity-related menstrual disorders or autoimmune premature ovarian insufficiency may uncover shared diagnostic signatures, leveraging insights from Tan et al (2024), who identified overlapping pathways between PCOS and metabolic-inflammatory conditions.

Recent studies[28] have systematically evaluated the potential application of anti-Müllerian hormone (AMH) in the diagnosis of PCOS, suggesting that AMH levels can reflect follicular reserve and ovarian responsiveness, offering certain diagnostic sensitivity. However, AMH is susceptible to multiple factors (such as age, body weight, and variations) in detection methods, making it insufficient for diagnosis when used alone. The key genes (S100A9, MYH9, and ATG5) identified in this study primarily characterize the immunometabolic features of PCOS from the perspectives of inflammatory responses and mitochondrial dysfunction. These mechanisms are complementary to the ovarian reserve status reflected by AMH. Combining the molecular diagnostic model developed in this study with ovarian function indicators (such as AMH) as a composite biomarker holds promise for enhancing the overall diagnostic accuracy of PCOS and subtype identification capabilities. Future studies may further explore the clinical utility of combined modeling involving AMH and immunometabolism-related genes, providing new strategies for the individualized diagnosis and stratified treatment of PCOS.

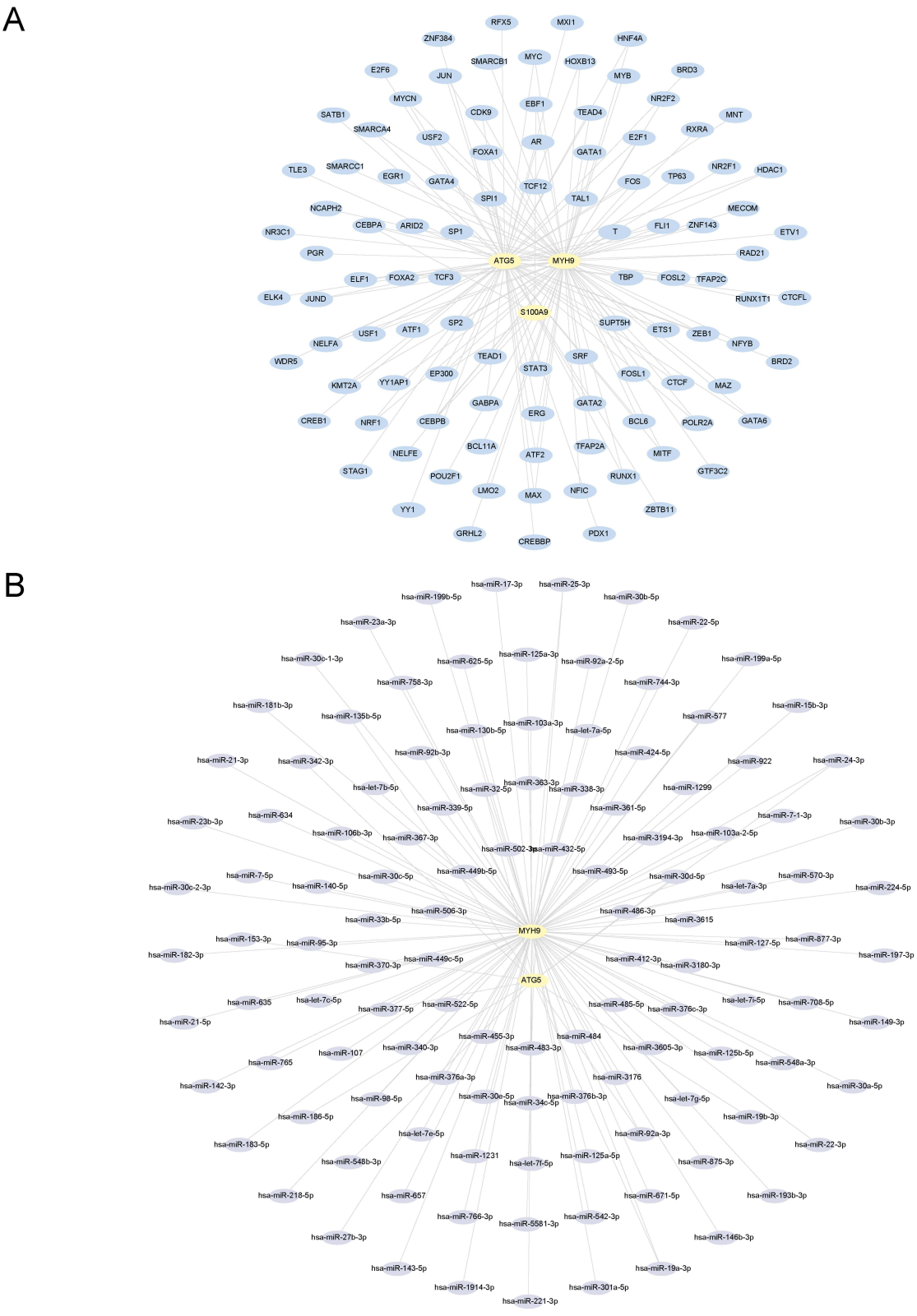

Figure 15 Regulatory Network of Key Genes. (**A**) mRNA-TF Regulatory Network of Key Genes. (**B**) mRNA-miRNA Regulatory Network of Key Genes. TF, Transcription Factor; Yellow is mRNA, blue is TF, and purple is miRNA.

https://doi.org/10.2147/IJWH.S560802

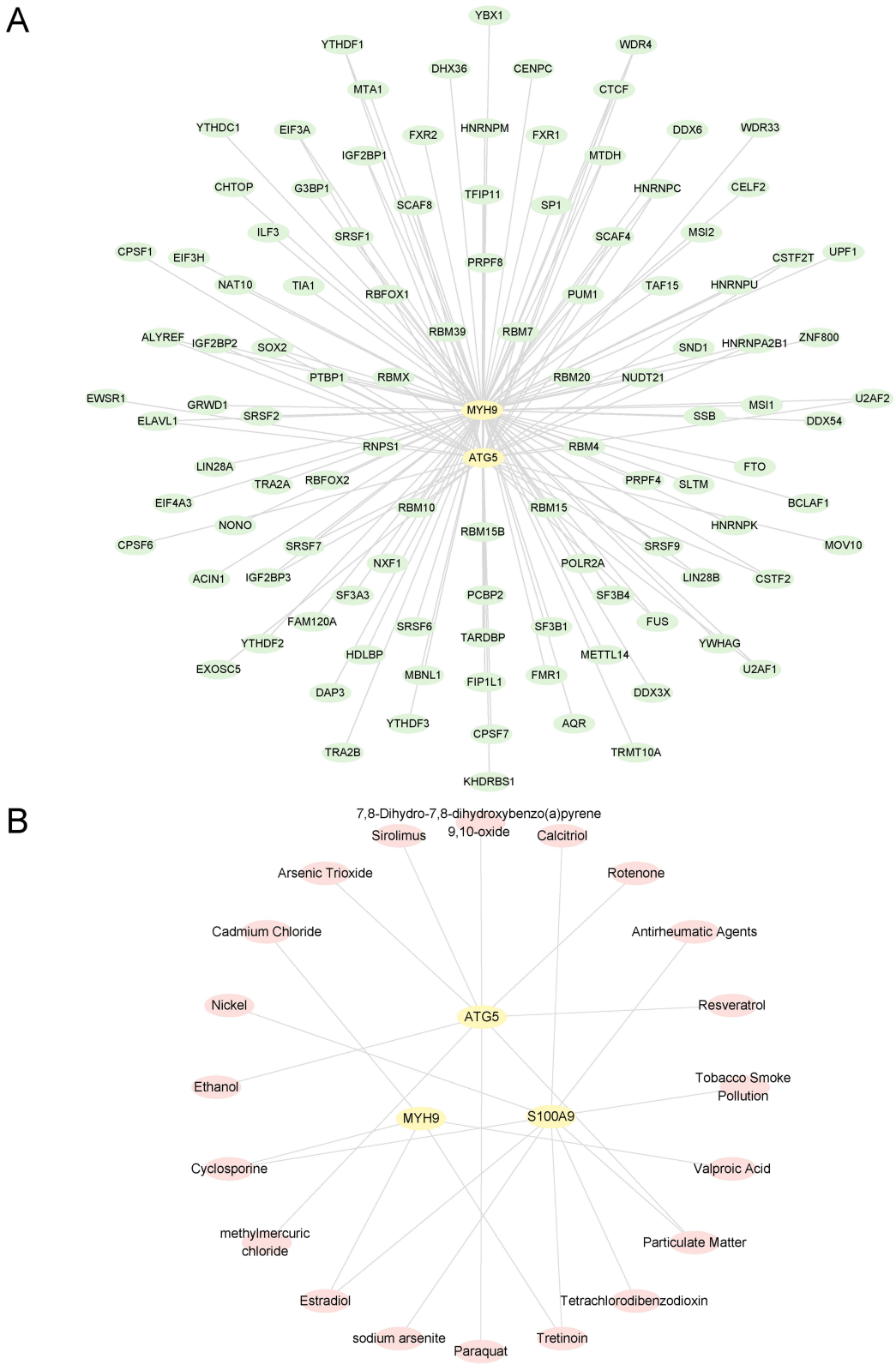

Figure 16 Regulatory Network of Key Genes. (**A**) mRNA-miRNA Regulatory Network of Key Genes. (**B**) mRNA-Drug Regulatory Network of Key Genes. TF, Transcription Factor; Yellow is mRNA, green is RBP, and pink is Drug.

Discussion on Immune Microenvironment: B Cell-CD8$^+$ T Cell Synergy

As indicated by immune infiltration analysis, activated B cells were strongly correlated with effector memory CD8$^+$ T cells (*r* = 0.968), alongside elevated mast cells and neutrophils in PCOS, suggesting a synergistic inflammatory axis. B cells may exacerbate chronic inflammation through antigen presentation and TNF-α release, while CD8$^+$ T cells contribute to ovarian tissue damage, as observed in PCOS granulosa cell apoptosis. Therapeutic strategies targeting this axis, such as B cell depletion or CD40L-CD40 signaling blockade, could disrupt inflammatory loops, while mast cell stabilizers may mitigate fibrosis and hyperandrogenism, as demonstrated in preclinical models. Notably, similar B cell/T cell interactions in lupus and rheumatoid arthritis suggest repurposing anti-B cell therapies for PCOS, offering a translational pathway for regulating immune-driven pathology.

Discussion on PPI: Functional Synergies

PPI network analysis confirmed interactions among ATG5, S100A9, and MYH9, with predicted roles in NLRP3 inflammasome activation and mitochondrial dynamics, suggesting a "NETs-mitophagy axis" where NETs impair mitochondrial clearance and exacerbate oxidative stress in PCOS. Mechanistic validation through co-immunoprecipitation (Co-IP) assays could confirm the interactions between S100A9 and mitochondrial proteins, while CRISPR-based ATG5 knockout in granulosa cells may clarify its role in NLRP3-driven apoptosis. Drug repurposing strategies, such as combining DNase I (NETs degradation) with mitophagy activators, could synergistically restore ovarian function, as proposed in preclinical models of metabolic-inflammatory dysregulation. The potential intervention molecules (DNase I and urolithin A) proposed in this study show promising application prospects in regulating PCOS-related inflammatory responses and mitochondrial functions. DNase I alleviates localized inflammation by degrading NETs and has demonstrated certain safety in studies related to autoimmune diseases and critical illnesses. However, excessive inhibition of NETs may weaken the host's anti-infection capacity. Hence, the benefits and risks of DNase I in PCOS still require careful evaluation. Urolithin A, a mitophagy activator derived from dietary polyphenol metabolites, has entered early-stage clinical research for metabolic diseases and exhibits acceptable overall tolerability. Nevertheless, direct evidence regarding its efficacy and safety in reproductive endocrine disorders is currently lacking. Future studies should conduct systematic animal experiments and prospective clinical trials in PCOS populations to verify the clinical feasibility and safety of the aforementioned targeted strategies. These findings highlight therapeutic opportunities for addressing the interconnected inflammatory and mitochondrial pathways in PCOS.

Strengths and Limitations

This study advanced the understanding of NETs and mitophagy in PCOS. However, there are some limitations. First, this study is limited by bioinformatics-driven conclusions requiring experimental validation. Second, Asian cohort dominance (78%) restricts the applicability of the results of this article, and cross-sectional data obscure temporal dynamics of NETs-mitophagy crosstalk. Therefore, future work must prioritize multi-ethnic cohorts and longitudinal analyses. This study holds significant importance in deepening the understanding of the mechanisms underlying the roles of NETs and mitophagy in PCOS, but it still has several limitations. First, although we have verified the expression changes of the key genes (S100A9, MYH9, and ATG5) in a PCOS mouse model, the overall results remain mainly based on bioinformatics analysis. Currently, there is a lack of experimental validation [such as quantitative real-time polymerase chain reaction (qRT-PCR) and immunohistochemistry (IHC)] in human ovarian tissues or granulosa cells, which to a certain extent limits the clinical translational potential of the conclusions. Second, the two GEO datasets used in this study (GSE34526 and GSE137684) are mainly derived from Asian women (India and China), resulting in a relatively homogeneous study population. Given the differences in genetic backgrounds, metabolic characteristics, and immune responses among different ethnic groups in PCOS, this population homogeneity may restrict the generalizability of the diagnostic model and immunometabolic findings of this study to non-Asian populations. Additionally, although batch effects from the public datasets were corrected using methods such as surrogate variable analysis (sva), the potential effects of residual batch differences and technical heterogeneity on some results cannot be completely ruled out. Therefore, subsequent work should involve experimental validation and longitudinal follow-up studies in human samples within multi-center,

multi-ethnic cohorts after obtaining ethical approval, to systematically evaluate the robustness and generalizability of the constructed diagnostic model and the NETs–mitophagy mechanism.

Conclusion

This study integrates multi-omics data to identify 18 NETs- and mitophagy-related genes (eg, S100A9 and ATG5) and constructs a high-accuracy diagnostic model (AUC > 0.9), advancing our understanding of PCOS pathogenesis. The proposed "NETs-mitophagy axis" highlights the interplay between chronic inflammation and mitochondrial dysfunction, offering actionable targets for therapeutic intervention. Physiologically, mitochondria maintain functional homeostasis via stringent quality control mechanisms. In PCOS, pathological stimuli including hyperandrogenism and insulin resistance directly disrupt mitochondrial integrity, triggering excessive reactive oxygen species (ROS) accumulation and aberrant mitochondrial DNA (mtDNA) release. These events activate downstream inflammatory cascades, particularly the NLRP3 inflammasome, which promote pro-inflammatory cytokine secretion and establish a persistent inflammatory microenvironment. This inflammatory state further exacerbates ovarian dysfunction and metabolic disturbances, forming a pathological feedback loop that perpetuates PCOS progression. The results revealed the synergistic mechanism of neutrophil extracellular traps (NETs)-mitochondrial dysfunction in the inflammatory response and metabolic abnormalities of polycystic ovary syndrome (PCOS). This study provides a theoretical basis for clarifying the immunometabolic pathological basis of PCOS and identifying novel molecular diagnostic targets, while offering a new research direction for the development of precision therapeutic strategies based on mitochondrial regulation. Future studies will focus on conducting external validation of the diagnostic model developed in this study and the proposed NETs-mitophagy axis in large-scale, multi-center, and multi-ethnic cohort. This will further evaluate their clinical applicability across diverse populations.

Data Sharing Statement

All study materials, including survey questionnaires, interview guides, and experimental protocols, are available from the corresponding author upon request. Researchers seeking access to the data must submit a formal request detailing the purpose of data use and agree to adhere to data protection protocols to ensure participant confidentiality.

Acknowledgments

The authors would like to express their sincere gratitude to all participants who volunteered to take part in this study, without whom the research would not have been possible.

Author Contributions

[Author Huang Ning]: Was responsible for coordinating the entire project, conceived and designed the research project, supervised the data collection process, led the data analysis, drafted the initial draft, and revised the academic content of the manuscript. Responded to peer review comments and ensured compliance with ethical standards. Performed bioinformatics and network pharmacology analyses. [Author Lou Luyun]: Participated in the research design, conducted field data collection, assisted in statistical analysis, and reviewed relevant literature. Corresponding Author: Provided theoretical guidance on pharmacology and network modeling; revised the manuscript critically; Supervised the study. All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

The design of this study, the collection, analysis and interpretation of the data, as well as the preparation and writing of the paper, all the expenses were borne by the authors of the paper.

Disclosure

All authors declare that they have no known competing financial or non-financial interests that could have influenced the design, conduct, analysis, or interpretation of the study. No author has received funding or other benefits from third parties that might be perceived as affecting the objectivity of this research. This includes, but is not limited to, employment relationships, consultancies, stock ownership, honoraria, and patent applications related to the study topic.

References

1. Davis SMP, Meltzer PS. GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor. *Bioinformatics*. 2007;23 (14):1846–1847. doi:10.1093/bioinformatics/btm254
2. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*. 2013;41(Database issue): D991–995. doi:10.1093/nar/gks1193
3. Kaur S, Archer KJ, Devi MG, Kriplani A, Strauss JF, Singh R. Differential gene expression in granulosa cells from polycystic ovary syndrome patients with and without insulin resistance: identification of susceptibility gene sets through network analysis. *J Clin Endocrinol Metab*. 2012;97 (10):E2016–2021. doi:10.1210/jc.2011-3441
4. Fishilevich S, Nudel R, Rappaport N, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database*. 2017;2017. doi:10.1093/database/bax028
5. Wu J, Zhang F, Zheng X, et al. Identification of renal ischemia reperfusion injury subtypes and predictive strategies for delayed graft function and graft survival based on neutrophil extracellular trap-related genes. *Front Immunol*. 2022;13:1047367. doi:10.3389/fimmu.2022.1047367
6. Teng ZH, Li WC, Li ZC, Wang YX, Han ZW, Zhang YP. Neutrophil extracellular traps-associated modification patterns depict the tumor microenvironment, precision immunotherapy, and prognosis of clear cell renal cell carcinoma. *Front Oncol*. 2022;12. doi:10.3389/fonc.2022.1094248
7. Wu L, Quan W, Luo Q, Pan Y, Peng D, Zhang G. Identification of an immune-related prognostic predictor in hepatocellular carcinoma. *Front Mol Biosci*. 2020;7:567950. doi:10.3389/fmolb.2020.567950
8. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882–883. doi:10.1093/bioinformatics/bts034
9. 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):e47. doi:10.1093/nar/gkv007
10. Ben salem K, Ben Abdelaziz A. Principal component analysis (PCA). *Tunis Med*. 2021;99(4):383–389.
11. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res*. 2019;47(D1):D419–d426. doi:10.1093/nar/gky1038
12. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27–30. doi:10.1093/nar/28.1.27
13. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics*. 2012;16(5):284–287. doi:10.1089/omi.2011.0118
14. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–15550. doi:10.1073/pnas.0506580102
15. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinf*. 2013;14(1). doi:10.1186/1471-2105-14-7
16. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739–1740. doi:10.1093/bioinformatics/btr260
17. Engebretsen S, Bohlin J. Statistical predictions with glmnet. *Clin Clin Epigenet*. 2019;11(1):123. doi:10.1186/s13148-019-0730-1
18. Sanz H, Valim C, Vegas E, Oller JM, Reverter F. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinf*. 2018;19(1):432. doi:10.1186/s12859-018-2451-4
19. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019;47(D1):D607–d613. doi:10.1093/nar/gky1131
20. Franz M, Rodriguez H, Lopes C, et al. GeneMANIA update 2018. *Nucleic Acids Res*. 2018;46(W1):W60–w64. doi:10.1093/nar/gky311
21. Zhou KR, Liu S, Sun WJ, et al. ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. *Nucleic Acids Res*. 2017;45(D1):D43–d50. doi:10.1093/nar/gkw965
22. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498–2504. doi:10.1101/gr.1239303
23. Vlachos IS, Paraskevopoulou MD, Karagkouni D, et al. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res*. 2015;43(Database issue):D153–159. doi:10.1093/nar/gku1215
24. Singh A. RNA-binding protein kinetics. *Nat Methods*. 2021;18(4):335. doi:10.1038/s41592-021-01122-6
25. Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res*. 2014;42(Database issue):D92–97. doi:10.1093/nar/gkt1248
26. Grondin CJ, Davis AP, Wiegers JA, et al. Predicting molecular mechanisms, pathways, and health outcomes induced by Juul e-cigarette aerosol chemicals using the Comparative Toxicogenomics Database. *Curr Res Toxicol*. 2021;2:272–281. doi:10.1016/j.crtox.2021.08.001
27. Xiao B, Liu L, Li A, et al. Identification and verification of immune-related gene prognostic signature based on ssGSEA for osteosarcoma. *Front Oncol*. 2020;10. doi:10.3389/fonc.2020.607622
28. Vale-Fernandes E, Pignatelli D, Monteiro MP. Should anti-Müllerian hormone be a diagnosis criterion for polycystic ovarysyndrome. An in-depth review of pros and cons. *Eur J Endocrinol*. 2025;192(4):R29–R43. doi:10.1093/ejendo/lvaf062