Observational Study

Medicine

OPEN

Identification of biomarkers for endometriosis based on summary-data-based Mendelian randomization and machine learning

Ziwei Xie, MS[a,b], Yuxin Feng, MS[a,b], Yue He, MS[a,b], Yingying Lin, MS[a,b], Xiaohong Wang, PhD[a,*]

Abstract

Endometriosis (EM) significantly impacts the quality of life, and its diagnosis currently relies on surgery, which carries risks and may miss early lesions. Noninvasive biomarkers are urgently needed for early diagnosis and personalized treatment. This study utilized the genome-wide association study dataset from FinnGen and performed Multi-marker Analysis of GenoMic Annotation (MAGMA) to identify genes significantly associated with EM. Differentially expressed genes (DEGs) were then analyzed, and an intersection selection was conducted to obtain the MAGMA-related DEGs. Gene Ontology and Kyoto Encyclopedia of Genes and Genomes enrichment analyses were performed to explore the biological functions of these genes. Summary-data-based Mendelian randomization was used to identify potential risk and protective genes. Subsequently, a machine learning model was used to further select key biomarkers. Single-cell RNA sequencing and consensus clustering were applied to analyze the expression of biomarkers and classify the EM samples into subgroups. Immune infiltration analysis was conducted to evaluate the molecular characteristics of these subgroups. MAGMA analysis identified 2832 genes significantly associated with EM, while 3055 DEGs were detected. Intersection analysis resulted in 437 MAGMA-related DEGs. Summary-data-based Mendelian randomization analysis identified 10 candidate genes, and after further selection using a machine learning model, three core biomarkers were validated: adenosine kinase, enoyl-CoA hydratase/3-hydroxyacyl CoA dehydrogenase, and CCR4-NOT transcription complex subunit 7. Single-cell RNA sequencing revealed the expression patterns of these biomarkers. Consensus clustering analysis classified 77 EM samples into two subgroups, with immune infiltration analysis showing significant differences in immune cell composition among the subgroups. This study successfully identified three core biomarkers for EM: adenosine kinase, enoyl-CoA hydratase/3-hydroxyacyl CoA dehydrogenase, and CCR4-NOT transcription complex subunit 7, which exhibit protective roles in EM.

Abbreviations: ADK = adenosine kinase, AUC = area under the curve, BTG1 = B-cell translocation protein 1, BP = biological process, Cis-eQTL = Cis-acting eQTL, CNOT7 = CCR4-NOT transcription complex subunit 7, DEGs = differentially expressed genes, EHHADH = enoyl-CoA hydratase/3-hydroxyacyl CoA dehydrogenase, EM = endometriosis, eQTL = expression quantitative trait loci, FUMA = functional mapping and annotation, GBM = gradient boosting machine, GWAS = genome-wide association study, IV = instrumental variable, KEGG = Kyoto Encyclopedia of Genes and Genomes, KNN = K-nearest neighbors, MAGMA = Multi-marker Analysis of GenoMic Annotation, MDGs = MAGMA-related differentially expressed gene, MR = Mendelian randomization, PCA = principal component analysis, ROC = receiver operating characteristic, scRNA-seq = single-cell RNA sequencing, SMR = summary-data-based Mendelian randomization, SNPs = single nucleotide polymorphism.

Keywords: biomarkers, endometriosis, machine learning, single-cell transcriptomics, summary-data-based Mendelian randomization

The authors declare that this research was funded by the National Administration of Traditional Chinese Medicine's High-Level Key Discipline Construction Project for Traditional Chinese Medicine – Clinical Integration of Traditional Chinese and Western Medicine (Grant No. zyyzdxk-2023104), the National Advantageous Traditional Chinese Medicine Specialty, the Gynecological Minimally Invasive and Integrated Pelvic Floor Clinical Research Center of Fujian Province (Grant No. X202202-Clinical Center), and the National Natural Science Foundation Project Basic Enhancement Plan Special Topic of Fujian Provincial People's Hospital (Grant No. JCZX202412).

The authors have no conflicts of interest to disclose.

The datasets generated during and/or analyzed during the current study are publicly available.

Ethical approval was not required for this study, as it only utilized publicly available, deidentified summary statistics, and all original studies had already obtained a specific ethical review and informed consent.

Supplemental Digital Content is available for this article.

[a] Department of Obstetrics and Gynecology, Affiliated People's Hospital of Fujian University of Traditional Chinese Medicine, Fujian, China, [b] First Clinical Medical College, Fujian University of Traditional Chinese Medicine, Fuzhou, China.

** Correspondence: Xiaohong Wang, Department of Obstetrics and Gynecology, Affiliated People's Hospital of Fujian University of Traditional Chinese Medicine, No. 602, 817 Middle Road, Taijiang District, Fuzhou City, Fujian Province, 350004, China (e-mail: wangxhsrm1972@163.com).*

How to cite this article: Xie Z, Feng Y, He Y, Lin Y, Wang X. Identification of biomarkers for endometriosis based on summary-data-based Mendelian randomization and machine learning. Medicine 2025;104:14(e41804).

Received: 15 January 2025 / Received in final form: 19 February 2025 / Accepted: 20 February 2025

http://dx.doi.org/10.1097/MD.0000000000041804

1. Introduction

Endometriosis (EM) is a chronic, inflammatory, and estrogen-dependent gynecological disease characterized by the presence of tissue resembling endometrium outside the uterine cavity, which may lead to infiltration.[1,2] Approximately 190 million women worldwide are affected by this condition, accounting for 10% to 15% of women of reproductive age.[3] The clinical manifestations of EM are diverse, including dysmenorrhea, irregular uterine bleeding, chronic pelvic pain, infertility, and other complications, severely impacting patients' quality of life.[4] Because of symptoms overlap with various other diseases, early diagnosis is often challenging. Currently, surgery remains the gold standard for diagnosis.[5] However, surgical procedures carry risks of complications, and early lesions may be missed during surgery.[6] Consequently, there is an imperative necessity to discover sensitive and specific noninvasive biomarkers for early detection.

In recent years, genetic research has provided new perspectives for the exploration of biomarkers. Studies have indicated that genetic and environmental factors are key contributors to the development of EM.[7] Research has shown a familial aggregation of EM[8–10] In a large study involving 3096 female twins in Australia, the estimated heritability was about 50%.[11] This high heritability suggests a significant genetic susceptibility to EM. Therefore, approaching from a genetic perspective may help identify new biomarkers, thereby advancing early diagnosis and treatment of the disease.

Genome-wide association studies (GWAS) are widely used techniques that have identified genetic variations associated with various complex diseases or traits through screening millions of single-nucleotide polymorphisms (SNPs).[12] Multi-marker Analysis of GenoMic Annotation (*MAGMA*) is a gene and gene set analysis tool based on GWAS data that can rapidly and flexibly associate SNPs discovered in GWAS with specific genes.[13] Summary-data-based Mendelian randomization (SMR) further integrates GWAS data with gene expression data from expression quantitative trait loci (eQTL) studies to prioritize potential pathogenic genes.[14] The SMR method extends the traditional concept of Mendelian randomization (MR), allowing examination of the hypothesized associations between genetically determined gene expression levels and disease phenotypes.

Machine learning techniques have become important tools for screening disease-related biomarkers because of their ability to automatically identify patterns in complex data.[15] In addition, the incorporation of machine learning techniques, especially in predictive modeling, adds a novel aspect to biomedical research.[16,17] These technologies aid in developing predictive models for EM, thereby improving diagnostic accuracy and supporting personalized treatment plans.

This study aimed to identify biomarkers for EM and construct a diagnostic model by integrating GWAS and RNA sequencing data, utilizing genetic and machine learning methods. The aim is to examine potential gene biomarkers linked to the genetics of EM. Figure 1 delineates the research technique utilized in this study.

2. Materials and methods

2.1. EM data sources

The EM GWAS data in this study were sourced from the FinnGen project (R12 version). FinnGen is a large-scale genomics project that analyzes over 500,000 samples from the Finnish Biobank, combining genetic variation with health data to enhance understanding of disease mechanisms and susceptibility. The project is a collaborative effort between Finnish research institutions, biobanks, and international industry partners.[18] The FinnGen dataset includes a total of 20,190 diagnosed cases and 130,160 control samples.

We obtained three messenger RNA (mRNA) datasets – GSE51981, GSE7305, and GSE25628 – from the Gene Expression Omnibus. GSE51981 is based on the GPL570 platform. Non-EM samples with other diseases were excluded, resulting in a final selection of 77 EM samples and 34 control samples for the training set. GSE7305, also based on the GPL570 platform, includes 10 normal samples and 10 EM samples; GSE25628 uses the GPL571 platform, selecting 6 normal samples and 7 EM samples, which together serve as the validation set. In addition, the single-cell RNA sequencing (scRNA-seq) dataset GSE179640 was sourced from the Gene Expression Omnibus database, from which 3 control samples and 12 EM samples were chosen for more in-depth single-cell level analysis.

2.2. Functional mapping and annotation of GWAS analysis of EM

Functional mapping and annotation (FUMA) is an online platform that integrates information from multiple sources to facilitate post-GWAS analyses, such as functional annotation and gene prioritization.[19] In SNP2GENE, default parameters were used to identify lead SNPs and candidate SNPs (e.g., setting the maximum *P* value for lead SNPs to <5e−8 and the maximum *P* value threshold for candidate SNPs to <.05). The r^2 threshold for defining independent significant SNPs was set to ≤0.6, while the second r^2 threshold for defining lead SNPs was set to ≤0.1. MAGMA analysis utilized the GTEx v8 database, which covers 54 tissue types and 30 common tissue types for gene expression analysis.

2.3. Identification and functional annotation of MAGMA-related differentially expressed genes

We identified differentially expressed genes (DEGs) in the GSE51981 dataset using the R package Limma,[20] with filtering criteria set at $|\log2FC| \geq 1$ and adjusted *P* value less than .05, as previously reported in our study.[21] Volcano plots were generated utilizing ggplot2 to visually depict upregulated and downregulated genes. Subsequently, DEGs were intersected with gene sets obtained from MAGMA to identify MAGMA-related differentially expressed genes (MDGs). For these MDGs, Gene Ontology, and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were conducted using the R package clusterProfiler[22] to explore related functional pathways, with a selection criterion set at *P* less than .05.

2.4. Candidate gene identification

Cis-acting eQTL (Cis-eQTL) summary data from the GTEx V8 dataset for whole blood and uterine tissues were obtained from the SMR website (https://yanglab.westlake.edu.cn/software/smr/#Overview).[23] This dataset, based on GTEx V8, includes cis-eQTL information for 49 human tissues, facilitating gene expression regulatory analysis. This study utilized the SMR tool (version 1.3.1)14 to focus on cis-eQTL data from whole blood and uterus, controlling for pleiotropic interference using the Heterogeneity In Dependent Instruments (HEIDI) test. When *P* is less than .01 and *P*-value for the HEIDI test greater than 0.05, the result indicates a significant association; if *P*-value for the HEIDI test is less than 0.05, pleiotropy is present. We determined potential risk or protective factors based on the sign of *B* values, with particular attention to genes with B greater than 0 and log2FC greater than or equal to 1, indicating a critical role in the pathological process, and thus potential risk factors. Conversely, genes with B less than 0 and log2FC less than or equal to −1 may exhibit protective effects, leading to the identification of potential candidate genes.

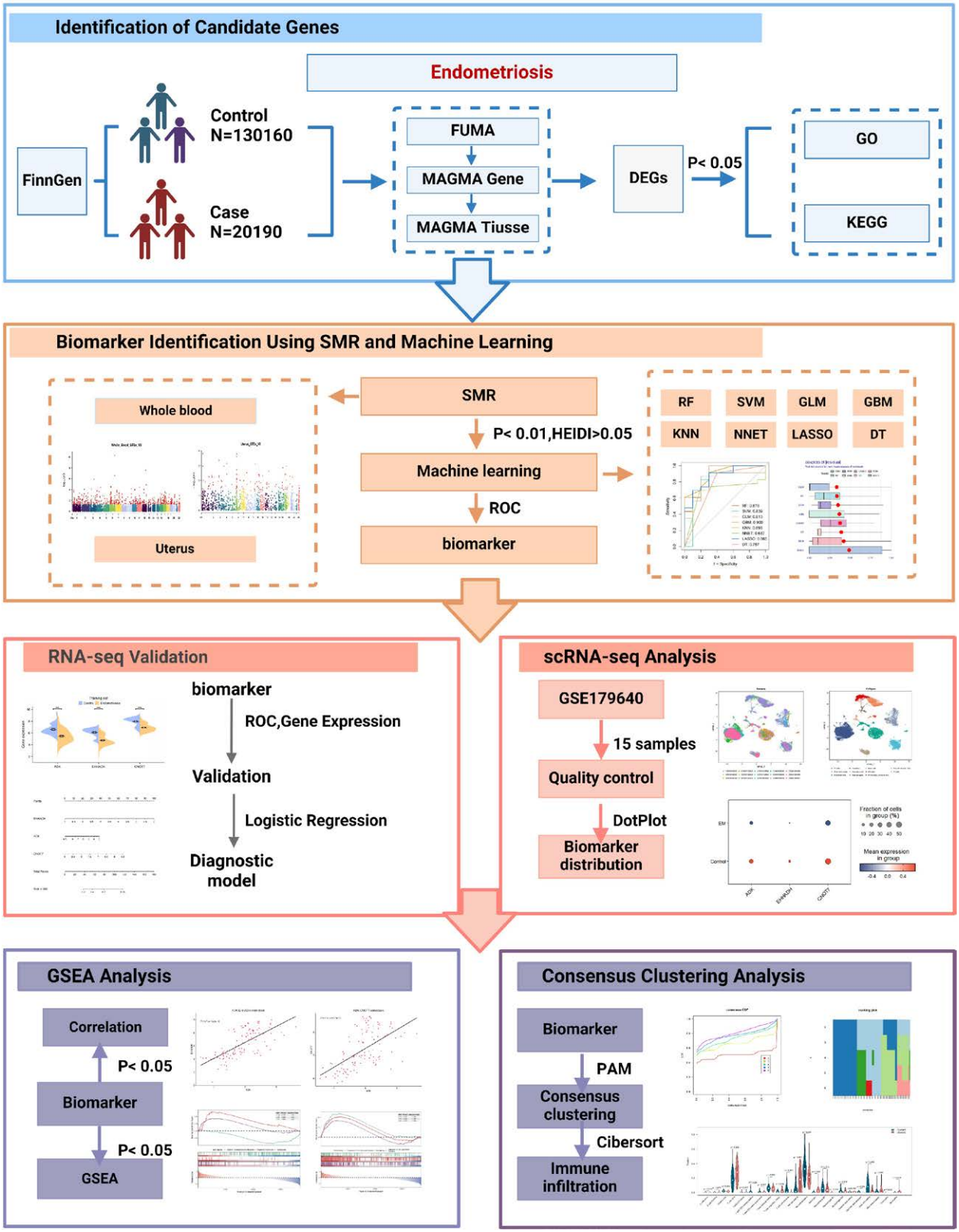

Figure 1. Flowchart of the research. DEGs = differentially expressed genes, DT = decision tree, EUR = European, FUMA = functional mapping and annotation, GBM = gradient boosting machine, GLM = generalized linear model, GO = Gene Ontology, GSEA = gene set enrichment analysis, HEIDI = Heterogeneity In Dependent Instruments, KEGG = Kyoto Encyclopedia of Genes and Genomes, KNN = K-nearest neighbors, LASSO = least absolute shrinkage and selection operator, NNET = neural network, PAM = partitioning around medoids, RF = random forest, ROC = receiver operating characteristic, scRNA-seq = single-cell RNA sequencing, SMR = summary-data-based Mendelian randomization, SVM = support vector machine.

2.5. Machine learning

We used 8 machine learning algorithms, including gradient boosting machine (GBM), generalized linear model, K-nearest neighbors (KNN), decision tree, random forest, least absolute shrinkage and selection operator regression, neural network, and support vector machine, to assess the importance of shared genes. To validate the stability and accuracy of the models, we utilized 5-fold repeated cross-validation, implemented through the R package caret. Subsequently, we analyzed each model using the R package DALEX, checking model performance through cumulative residual distribution and employing boxplots and receiver operating characteristic (ROC) curves with area under the curve (AUC) metrics to select the best-performing model.

2.6. Mendelian randomization analysis

In this study, cis-eQTL data for 3 candidate genes were obtained from GTEx V8 whole blood samples. The reliability of MR causal estimates is based on three core assumptions: the genetic variants used as instrumental variables (IVs), such as SNPs, must significantly and robustly predict the exposure; IVs must only be associated with the outcome through the exposure and have no direct effects; and IVs must not be associated with any potential confounders that could influence the exposure–outcome relationship. A two-sample MR approach was used using the TwoSampleMR software package,[24] with genes as exposures and diseases as outcomes. Genetic variants from multiple regions of the genome were selected as IVs. SNPs significantly associated with the exposure ($P < 5 \times 10^{-8}$) were screened, and linkage disequilibrium thresholds were set ($r^2 = 0.1$, kb = 10,000) to ensure independence. The strength of each SNP as an instrument was assessed using the *F*-statistic, with *F* greater than 10 considered indicative of a strong instrument. For genes with only 1 IV, the Wald ratio method was applied; for genes with multiple IVs, the inverse variance weighted method was used, with a *P* value below .05 considered evidence of a significant causal relationship. Heterogeneity was assessed using Cochran's *Q* test, with a *P* value less than .05 indicating significant heterogeneity. Horizontal pleiotropy was evaluated using the intercept from MR-Egger regression, where a *P* value less than .05 suggested evidence of pleiotropy. Only genes that passed validation and met the above criteria were considered for further analysis.

2.7. Validation of diagnostic models and biomarkers

On the basis of the expression of biomarkers, a diagnostic model for patients with EM was developed using the R package rms.[25] We plotted ROC curves using the R package pROC.[26] Calibration curves were then constructed to evaluate the predictive accuracy and reliability of the model. To obtain a larger sample size, we removed batch effects from GSE7305 and GSE25628 using the R package SVA,[27] creating a larger validation set. Finally, we validated the performance of the EM diagnostic model using ROC curves from the validation set and assessed the diagnostic capability of biomarkers in both the test and validation sets, considering genes with an AUC value greater than 0.7 as reliable biomarkers.[28] We subsequently validated the expression levels of biomarkers in both the test set and the test dataset.

2.8. Single-cell analysis

We analyzed scRNA-seq data using the R package Seurat.[29] During the quality control process, we removed low-quality data by excluding cells with mitochondrial gene expression levels exceeding 25%, as well as those with detected gene counts below 200 or above 10,000. In addition, we excluded cells with blood hemoglobin gene expression ratios exceeding 3% to avoid interference from blood-related cells. To further ensure data quality, we also removed outlier cells with total transcript counts (nCount RNA) exceeding 100,000. After quality control, normalization, identification of highly variable genes, scaling, and principal component analysis (PCA) were performed on the remaining cells. To address batch effect issues from different datasets, we used the R package Harmony[30] to eliminate these effects. Finally, different cell subclusters were identified using the FindNeighbors and FindClusters functions. To ensure accuracy in cell labeling, we utilized the FindAllMarkers function to detect DEGs within each cluster, conducting a comprehensive assessment of final cell types using predefined markers from existing literature.

2.9. Correlation and gene set enrichment analysis of biomarkers

We determined the correlation between biomarkers using the cor function. In addition, we performed gene set enrichment analysis on each potential biomarker in the EM dataset using the R package clusterProfiler.[22] Gene sets were obtained from the MSigDB[31] database, specifically the c2.cp.kegg_legacy.v2024.1.Hs.entrez.gmt and c5.go.bp.v2024.1.Hs.entrez.gmt collections. Adjusted *P* values below .05 were deemed statistically significant.

2.10. Subgroup analysis and immune infiltration assessment

Consensus clustering is a commonly used computational method to determine the optimal number of unsupervised clusters within a dataset. In this study, we utilized the R package ConsensusClusterPlus[32] to perform clustering analysis on the 77 EM samples, dividing them into multiple clusters. To select the optimal number of clusters, we used various analytical methods, including consistency matrix plots, cumulative distribution function plots, assessments of the relative change in cumulative distribution function area, and silhouette plots. To further explore the characteristics of disease subtypes, PCA was conducted on the clustering results, visually illustrating the similarities and differences between clusters. CIBERSORT is a robust tool for inferring the cellular composition of complex tissues based on gene expression profiles.[33] We utilized the R package CIBERSORT to evaluate the immune infiltration levels in each disease subtype and identify significantly different immune cell types.

2.11. Statistical analysis

All bioinformatics analyses were conducted using R software. Disparities between two groups were evaluated via the Wilcoxon test. Statistical significance was determined at *P* less than .05, with the subsequent designations: $P < .05$ as *, $P < .01$ as **, and $P < .001$ as ***.

3. Results

3.1. FUMA analysis

We utilized the FUMA platform for FUMA of the EM data, analyzing the distribution of SNPs across the genome. The SNP2GENE analysis of linkage disequilibrium regions identified a total of 5831 candidate SNPs, 276 independent significant SNPs, 90 lead SNPs, and 32 genomic risk loci (Supplementary Table S1, Supplemental Digital Content, https://links.lww.com/MD/O629). In addition, through MAGMA analysis, we identified 2832 genes significantly associated with the trait (Supplementary Table S2, Supplemental Digital Content,

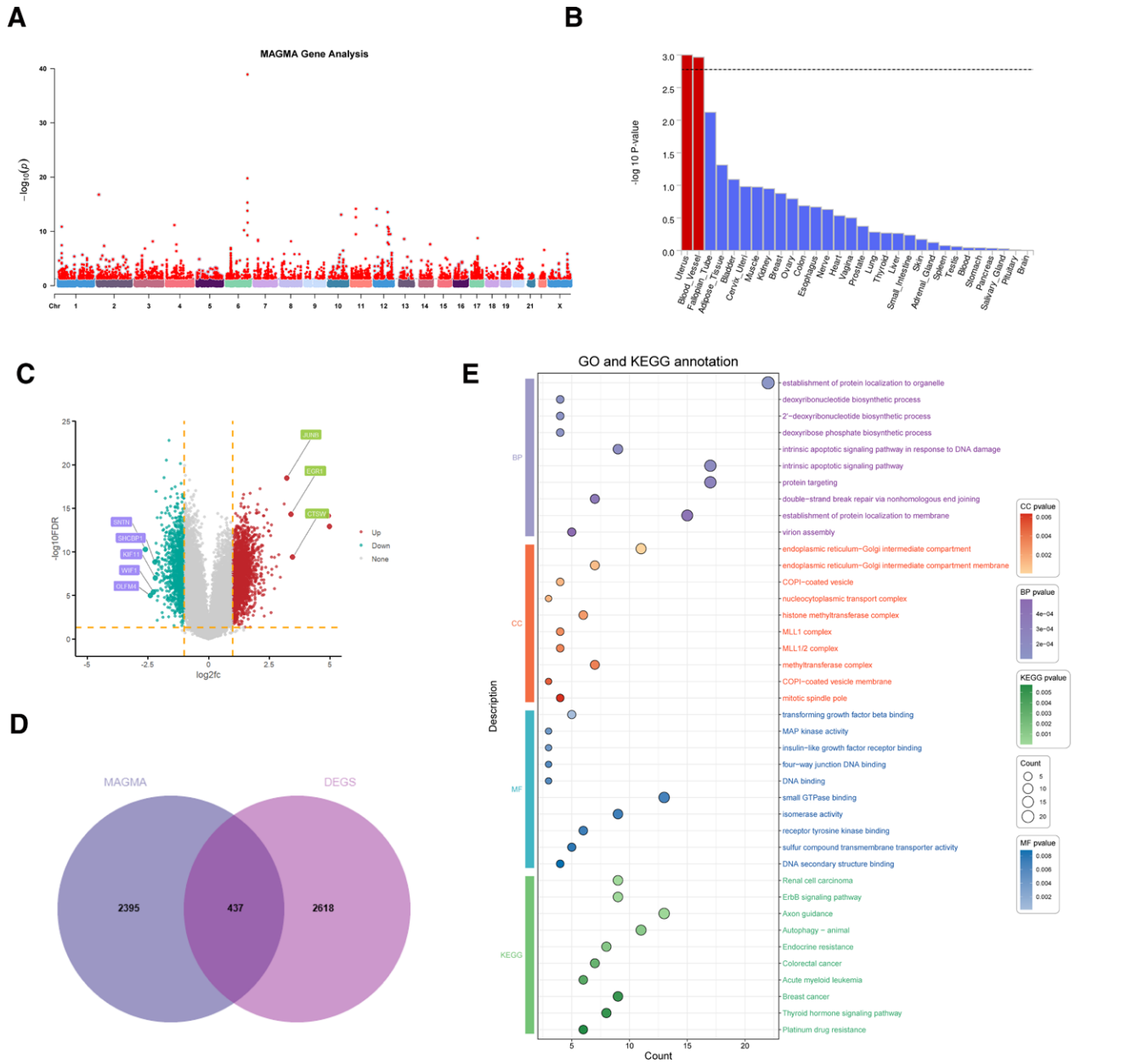

Figure 2. Identification and enrichment analysis of MDGs. (A) The gene Manhattan map depicts the distribution and significance of genes linked to the trait of interest across several chromosomes. (B) The MAGMA study findings demonstrate gene expression levels in several organs, utilizing data from the GTEx v8 database. (C) Volcano plot of differential gene expression. (D) Venn diagram of overlapping genes between DEGs and MAGMA genes. (E) Functional enrichment analysis of MDGs. DEG = differentially expressed gene, GO = Gene Ontology, KEGG = Kyoto Encyclopedia of Genes and Genomes, MAGMA = Multi-marker Analysis of GenoMic Annotation, MDG = MAGMA-related differentially expressed gene.

https://links.lww.com/MD/O630 and Fig. 2A). These genes may play important roles in the pathogenesis of EM and exhibit tissue-specific expression patterns. Notably, these genes are highly expressed in the uterus and blood vessels (Fig. 2B), suggesting their involvement in biological processes (BPs) related to EM.

3.2. MDGs identification and enrichment analysis

From the GSE51981 dataset, a total of 3055 DEGs were identified, including 1922 upregulated genes and 1133 downregulated genes. A volcano plot clearly illustrates the distribution of these DEGs (Fig. 2C). Further intersection analysis with MAGMA genes led to the identification of 437 MDGs (Fig. 2D). Subsequently, we performed Gene Ontology and KEGG enrichment analyses to further assess the potential biological functions of these MDGs (Fig. 2E, Supplementary Tables S3 and S4, Supplemental Digital Content, https://links.lww.com/MD/O631). According to the results of KEGG pathway enrichment analysis, the MDGs were significantly enriched in multiple pathways associated with cell proliferation, metabolism, and signaling, such as the phosphoinositide 3-kinase–protein Kinase B signaling pathway, hypoxia-inducible factor 1 signaling pathway, and vascular endothelial growth factor signaling pathway. These pathways play crucial roles in cell growth, survival, angiogenesis, and energy metabolism, suggesting

that EM may be closely linked to abnormal cell proliferation and angiogenesis. In terms of BPs, the DEGs were mainly involved in protein localization, apoptosis signaling, DNA damage repair, cell proliferation regulation, and metabolic processes such as nucleotide metabolism and glucose metabolism. These results indicate that EM may be closely associated with abnormal cell proliferation, apoptosis regulation, and metabolic disorders. In cellular components, the MDGs were significantly enriched in organelles or structures such as the endoplasmic reticulum–Golgi intermediate compartment, coat protein complex I-coated vesicles, and cell-matrix junctions, suggesting that EM may involve intracellular protein transport, cytoskeleton organization, and extracellular matrix interactions. In molecular functions, the MDGs were primarily involved in DNA binding, kinase activity, GTPase binding, and cytokine binding, indicating that EM may be related to abnormal signal transduction, gene expression regulation, and intercellular communication. These findings provide new clues for a deeper understanding of the molecular mechanisms underlying endometriosis.

3.3. Candidate genes identified by SMR

Through SMR analysis, using a screening criterion of *P* value less than .01 and HEIDI > 0.05, we identified genes in both whole blood and uterine tissue (Fig. 3A and B, Supplementary Tables S5 And S6, Supplemental Digital Content, https://links.lww.com/MD/O632). In whole blood samples, 156 significantly associated genes were identified, while 30 associated genes were found in uterine tissue. On the basis of the expression trends of the MDGs, we ultimately selected 10 candidate genes with potential biological significance (Fig. 3C). Among these, 4 genes were significantly upregulated, potentially serving as risk factors, while 6 genes were significantly downregulated, suggesting a protective role.

3.4. Candidate genes identified by machine learning

We performed machine learning analysis on the 10 candidate genes to assess the classification performance of various models. The results showed that the GBM model had the smallest

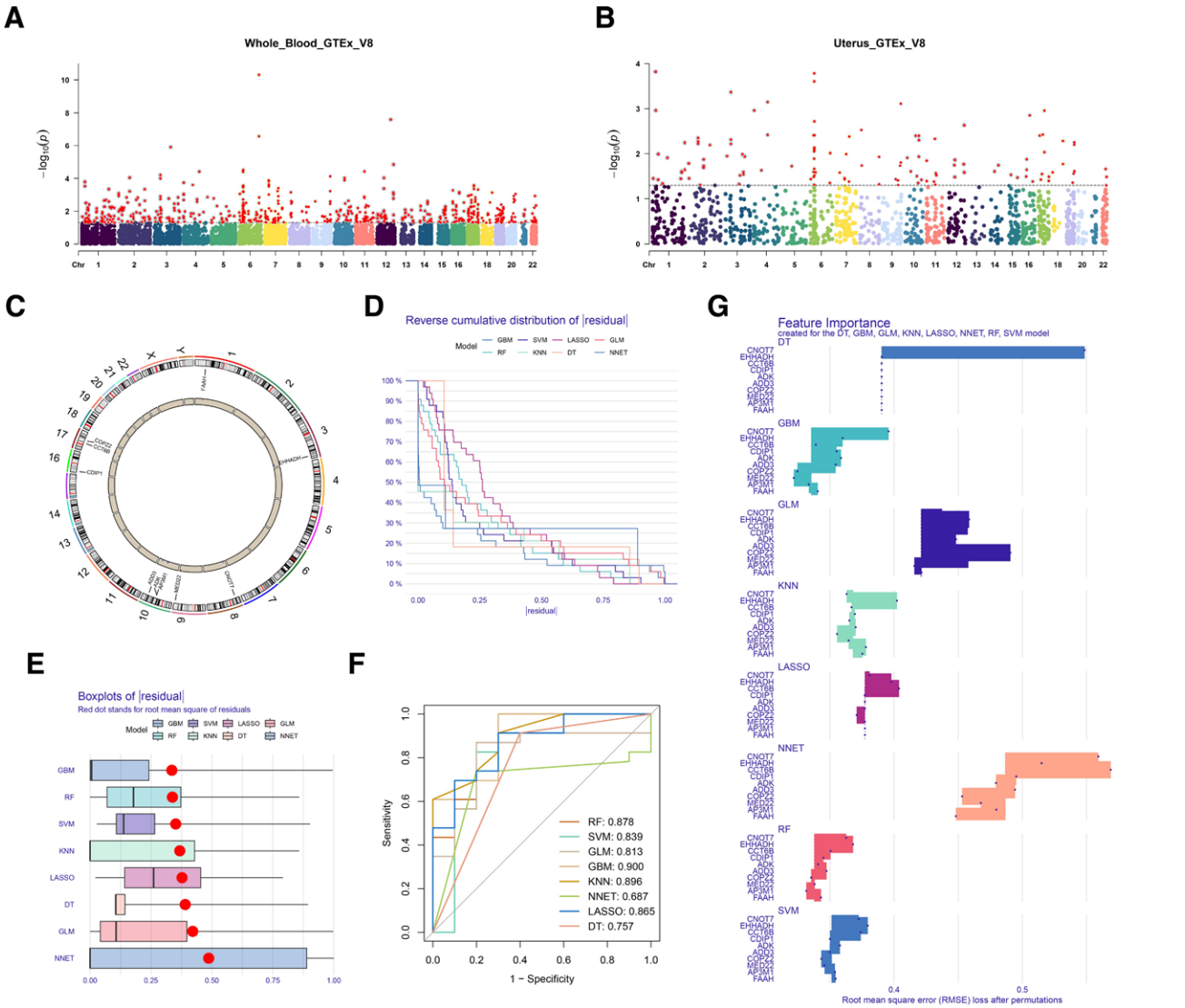

Figure 3. SMR and machine learning for identifying biomarkers. (A) SMR analysis results in whole blood. (B) SMR analysis results in uterus. (C) Circular plot illustrating the MR results for the analyzed genes. (D) Reverse cumulative distribution of absolute residuals across machine learning models. (E) Boxplots of absolute residuals across machine learning models. (F) ROC curves for machine learning models. (G) Feature importance across machine learning models. DT = decision tree, GBM = gradient boosting machine, GLM = generalized linear model, KNN = K-nearest neighbors, LASSO = least absolute shrinkage and selection operator, NNET = neural network, RF = random forest, MR = Mendelian randomization, RF = random forest, ROC = receiver operating characteristic, SMR = summary-data-based Mendelian randomization, SVM = support vector machine.

residuals (Fig. 3D and E) and achieved the highest AUC value among all models. Specifically, the AUC values for each model were as follows: GBM (0.900), KNN (0.896), random forest (0.878), least absolute shrinkage and selection operator (0.865), support vector machine (0.839), generalized linear model (0.813), decision tree (0.757), and neural network (0.687) (Fig. 3F). This suggests that the GBM model has a stronger ability to differentiate between patients in different clusters. The top 10 most significant variables in each model were selected based on the root mean square error. The top three key predictive factors selected from the GBM model (adenosine kinase [*ADK*], enoyl-CoA hydratase/3-hydroxyacyl CoA dehydrogenase [*EHHADH*], and CCR4-NOT transcription complex subunit 7 [*CNOT7*]) were used for subsequent analysis (Fig. 3G).

3.5. Verification of candidate genes

We successfully analyzed the causal relationships between *ADK*, *EHHADH*, and *CNOT7* and EM. The results showed significant causal associations between these three genes and EM (Supplementary Table S7, Supplemental Digital Content, https://links.lww.com/MD/O633). Sensitivity analysis revealed minimal evidence of horizontal pleiotropy, indicating that the results were not significantly influenced by bias. Moreover, the *F* statistics for the genetic instruments associated with each gene were high, demonstrating the strength and reliability of the IVs, further confirming the robustness of the causal inferences. These findings further support their potential as biomarkers.

3.6. Validation of diagnostic model and biomarkers

To visually demonstrate the performance of the diagnostic model, we constructed a risk nomogram for EM incorporating three biomarkers (Fig. 4A). The calibration curve indicates that the model has good predictive value (Fig. 4B). Decision curve analysis further indicated that when the threshold probability for patients or clinicians exceeds 10%, using the diagnostic model to predict the occurrence of EM yields greater net benefit compared with the "diagnose all patients" or "diagnose no patients" scenarios (Fig. 4C). Moreover, the AUC values of the diagnostic model in the training and test sets were 0.9 and 0.95, respectively, suggesting that the model's discriminatory ability is stable across different datasets (Fig. 4D).

We further assessed the diagnostic ability of the three biomarkers (*ADK*, *EHHADH*, and *CNOT7*) by evaluating ROC curves in both the training and test sets (Fig. 4E). The results

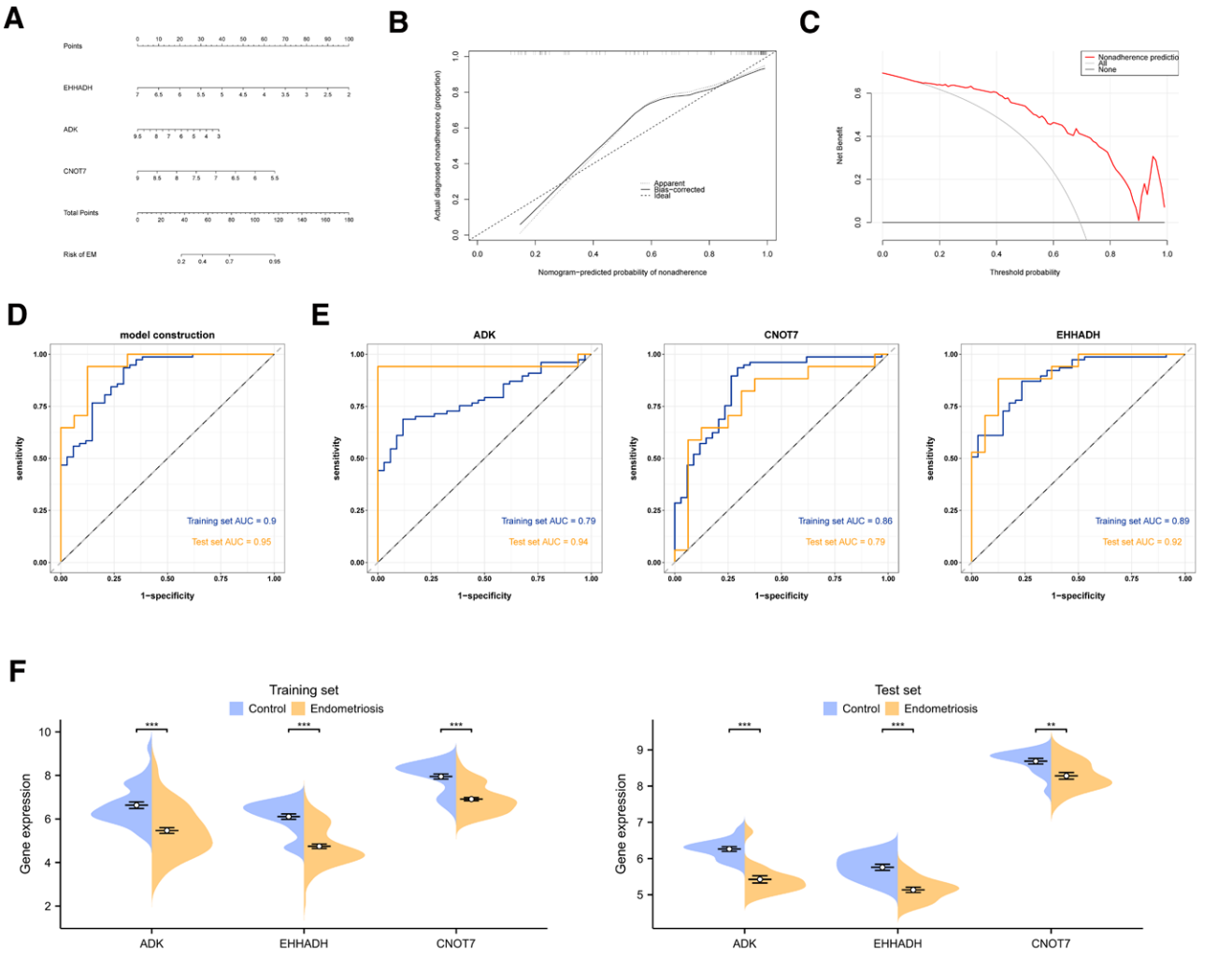

Figure 4. Validation of diagnostic models and biomarkers. (A) Nomogram for predicting risk of EM. (B) Calibration plot for nomogram-predicted probability of nonadherence. (C) The decision curve graph evaluates the net benefit of diagnostic predictions at various threshold probabilities. (D) ROC curves for diagnostic model on training and test sets. (E) ROC curves for biomarker on training and test sets. (F) Gene expression levels of biomarker in control and EM groups. *ADK* = adenosine kinase, AUC = area under the curve, *CNOT7* = CCR4-NOT transcription complex subunit 7, *EHHADH* = enoyl-CoA hydratase/3-hydroxyacyl CoA dehydrogenase, EM = endometriosis, ROC = receiver operating characteristic.

showed that the AUC for *ADK* was 0.79 in the training set and 0.94 in the test set; for CNOT7, it was 0.86 in the training set and 0.79 in the test set; and for *EHHADH*, it was 0.89 in the training set and 0.92 in the test set. These results suggest that all biomarkers possess good diagnostic ability. In addition, we analyzed the gene expression differences of these three independent biomarkers in both the training and test sets (Fig. 4E). The results showed that the expression of *ADK*, *EHHADH*, and *CNOT7* was significantly reduced in the EM group, and all differences were highly statistically significant. In the test set, the expression trend of the biomarkers was consistent with that in the training set and also showed statistical significance. These consistent expression trends and strong diagnostic performance further support the potential value of *ADK*, *EHHADH*, and *CNOT7* as diagnostic biomarkers.

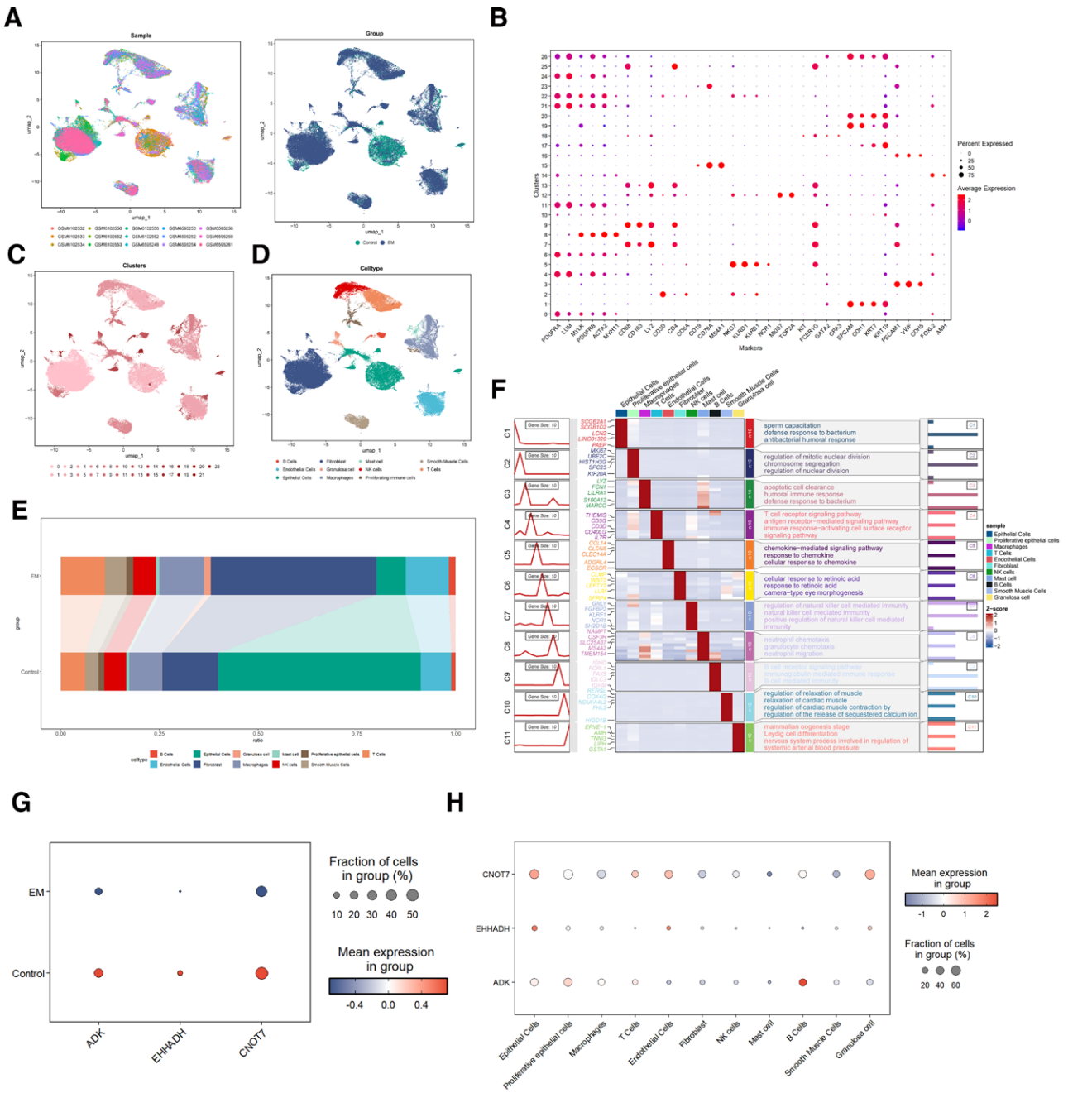

Figure 5. scRNA-seq analysis. (A) The UMAP map depicts the distribution of cells, with colors representing the source of samples (control and EM samples). (B) The DotPlot illustrates the expression levels of established cell-type marker genes across several cell clusters, with colors indicating average expression and dot size denoting the proportion of cells expressing the gene. (C) Decision curve analysis for nomogram-predicted nonadherence. (D) The UMAP plot shows that the cells are identified as 11 different cell types. (E) Proportional variations in the 11 cell types between normal and EM samples are presented. (F) The left panel illustrates the dynamic patterns of representative DEGs; the middle panel showcases a heatmap of DEGs across clusters; the right panel exhibits the results of GO enrichment analysis for each cluster. (G) Gene expression levels of biomarker in control and EM groups. (H) Gene expression levels of biomarker in 11 different cell types. *ADK* = adenosine kinase, *CNOT7* = CCR4-NOT transcription complex subunit 7, DEG = differentially expressed gene, *EHHADH* = enoyl-CoA hydratase/3-hydroxyacyl CoA dehydrogenase, EM = endometriosis, GO = Gene Ontology, scRNA-seq = single-cell RNA sequencing, UMAP = Uniform Manifold Approximation and Projection.

3.7. Single-cell analysis

After strict quality control and screening, a total of 67,208 cells were retained for further analysis, including 52,243 cells from EM samples and 14,965 cells from control samples (Fig. 5A). We performed dimensionality reduction and clustering analysis using Uniform Manifold Approximation and Projection, successfully identifying 23 distinct cell subpopulations. We analyzed the expression of known lineage markers for these 23 cell subpopulations in both the normal and EM groups (Fig. 5B and C). The results revealed that we identified 11 major cell types, including B cells, endothelial cells, epithelial cells, fibroblasts, granulocytes, macrophages, mast cells, natural killer cells, proliferative immune cells, smooth muscle cells, and T cells (Fig. 5D). By comparing the cell composition distribution between the normal and EM groups, we found that fibroblasts and granulocytes were significantly increased in the EM group, while epithelial cells were decreased. Given the close relationship between EM and the immune system, we specifically investigated changes in immune cells and found that T cells and macrophages were increased in the EM group (Fig. 5E).

In addition, we performed BP enrichment analysis on the top 10 highly expressed genes in each cell subpopulation, further revealing the potential mechanisms of EM (Fig. 5F). In the biomarker distribution analysis, we observed that the average expression levels of *ADK*, *EHHADH*, and *CNOT7* were lower in the EM group compared with the control group (Fig. 5G), consistent with the RNA sequence results and further supporting their potential as biomarkers. Notably, *ADK* was primarily expressed in B cells, *EHHADH* was concentrated in epithelial cells, and *CNOT7* was distributed in epithelial cells, granulocytes, and endothelial cells without distinct specificity (Fig. 5H).

3.8. Functional exploration of biomarkers

In gene correlation analysis, we found that *ADK*, *EHHADH*, and *CNOT7* were positively correlated with each other, and all results were statistically significant (Fig. 6A, Supplementary Table S8, Supplemental Digital Content, https://links.lww.com/MD/O634). Furthermore, GSEA revealed that the roles of *ADK*, *EHHADH*, and *CNOT7* in key pathways and BPs associated with EM varied, but they worked synergistically. In the KEGG pathways, it is noteworthy that *ADK*, *EHHADH*,

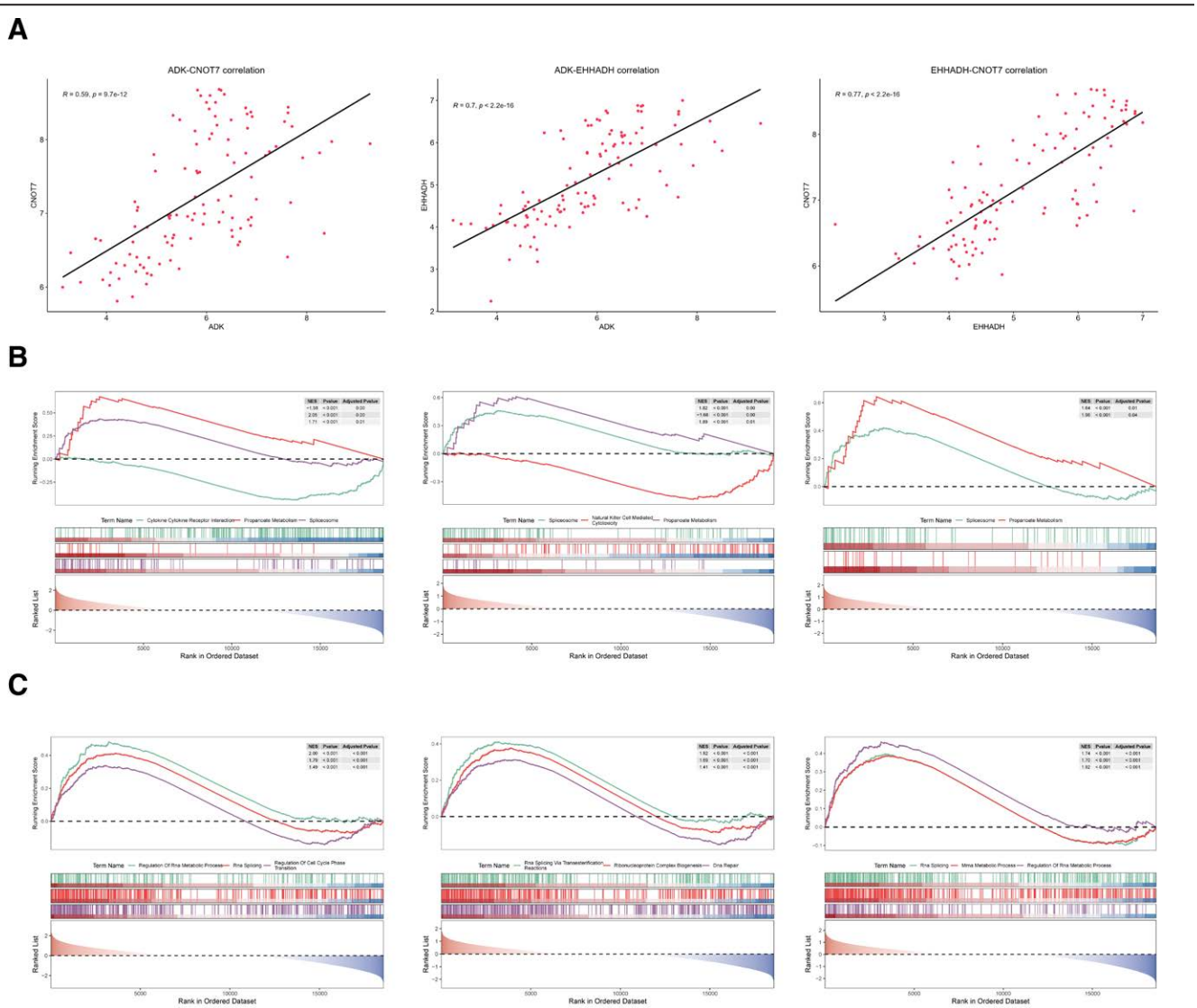

Figure 6. Correlation and GSEA analysis of biomarkers. (A) Correlation analysis of gene expression levels between *ADK*, *EHHADH*, and *CNOT7*. (B and C) Gene set enrichment analysis results of biomarkers. *ADK* = adenosine kinase, *CNOT7* = CCR4-NOT transcription complex subunit 7, GSEA = gene set enrichment analysis.

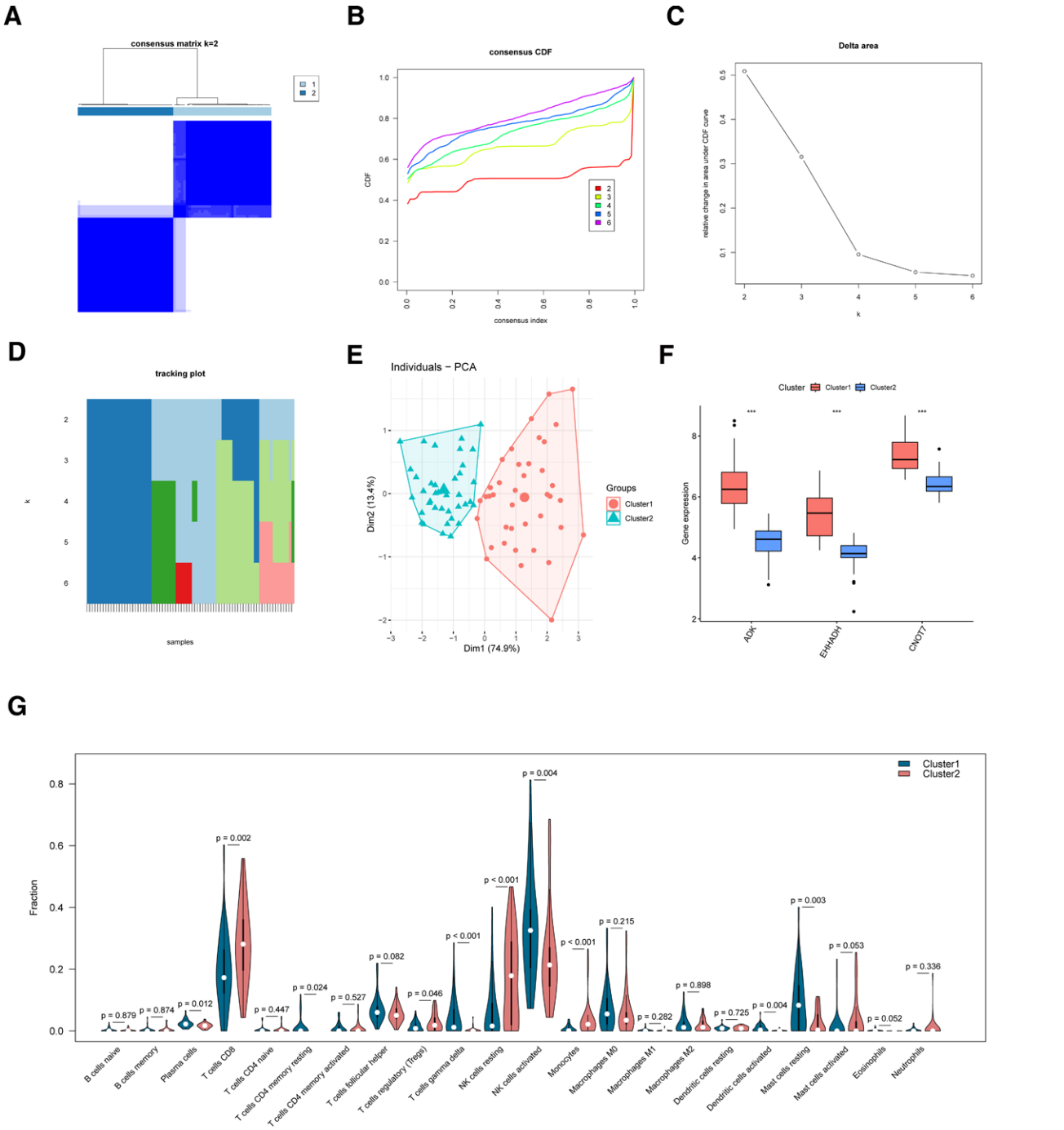

Figure 7. Consensus clustering analysis. (A) Consensus clustering matrix ($k = 2$) based on three biomarkers. (B) Consensus CDF plot for cluster stability. (C) Delta area plot for determining optimal cluster number (k). (D) Tracking plot for cluster membership across different values of k. (E) PCA plot of two clusters. (F) Gene expression levels of three biomarkers across two clusters. (G) Violin plot of immune cell fractions across two clusters. CDF = cumulative distribution function, PCA = principal component analysis.

and *CNOT7* were significantly enriched in pathways such as "SPLICEOSOME" and "PROPANOATE_METABOLISM," suggesting that they may drive the development of EM through the regulation of energy metabolism and RNA splicing abnormalities. In terms of BP, *ADK*, *CNOT7*, and *EHHADH*, as key regulators of EM, were significantly enriched in pathways related to RNA splicing, mRNA metabolism, and ribonucleoprotein complex biogenesis, indicating that they may lead to widespread gene expression dysregulation by interfering with posttranscriptional regulatory networks (Fig. 6B and C). The synergistic roles of these genes in multiple pathways and BPs provide new insights into the molecular mechanisms underlying EM and offer a potential foundation for targeted therapeutic strategies for the disease.

3.9. Consistency clustering identifies two subclusters

In this study, to explore the potential impact of different gene expression patterns on EM samples, we used unsupervised hierarchical clustering based on the expression characteristics of the three biomarkers to group 77 samples. The optimal clustering result was achieved when setting $k = 2$, dividing the samples

into two major subclusters (Fig. 7A–D, Supplementary Table S9, Supplemental Digital Content, https://links.lww.com/MD/O635). This grouping result was further validated by PCA, which showed significant differences in gene expression patterns between the different subclusters (Fig. 7E). We also generated box plots to display the expression levels of each gene in the different subclusters, highlighting the significant intergroup differences (Fig. 7F).

Subsequently, the immune infiltration results from CIBERSORT indicated significant differences in the proportion of immune cells between the two subclusters. Specifically, the proportions of plasmacytes, CD8 T cells, memory resting CD4 T cells, regulatory T cells, γδ T cells, resting natural killer cells, activated natural killer cells, monocytes, activated dendritic cells, and resting mast cells were markedly different (Fig. 7G). These findings suggest that different gene expression patterns are closely associated with the immune characteristics of EM samples, providing valuable insights for further investigating the molecular mechanisms of EM and precision stratification.

4. Discussion

EM imposes significant social costs, negatively affecting the quality of life for women and their families. Although the exact etiology and pathogenesis of EM are still unclear, existing studies suggest a significant association with genetic factors. In-depth research into genetic factors will not only help understand the underlying causes of the disease but also aid in identifying more effective early biomarkers.

This study used genetic and bioinformatics approaches, based on MAGMA and differential expression analysis, to identify 437 DEGs associated with EM. Enrichment analysis revealed that the MDGs are closely related to multiple cancer-related cell signaling pathways (such as phosphoinositide 3-kinase–protein Kinase B signaling pathway, Notch, ErbB, etc.), suggesting that the disease may involve abnormalities in cell proliferation and metastasis processes. Furthermore, endocrine resistance, hormone signaling pathways, and metabolic disorders may play an important role in the onset and progression of EM, especially in treatment response and disease resistance. Immune escape mechanisms, including the programmed death-ligand 1/programmed death-1 checkpoint pathway and its relationship with virus-induced carcinogenesis, may also be important factors in the development of EM.

Through SMR analysis, we ultimately identified 10 candidate genes. MR and systematic machine learning analysis highlighted three core biomarkers – *ADK*, *EHHADH*, and *CNOT7* – as potentially protective in EM. scRNA-seq further validated the potential of these biomarkers and revealed their expression and distribution characteristics within cells, confirming their potential as biomarkers for EM.

ADK encodes adenosine kinase, which belongs to the ribokinase family. It indirectly regulates extracellular adenosine levels by phosphorylating intracellular adenosine to *5′*-adenosine monophosphate,[34] making it a key regulator of adenosine. In our study, *ADK* was identified as a protective factor, although its specific association with EM remains unclear. ADK dysfunction is associated with various pathologies, including diabetes, epilepsy, and cancer, and thus, *ADK* is also considered a potential therapeutic target. Under normal conditions, the balance between adenosine and ADK is tightly maintained. Once *ADK* expression is altered, it can affect adenosine receptor activation and play a key role in various pathological processes.[35] Physiologically, the main function of adenosine is to protect tissues and resist damage, balancing the proimmunogenic and proinflammatory effects of eATP. However, under pathological conditions, elevated adenosine levels are closely associated with anti-inflammatory effects in tissues and inhibitory antitumor immunity in various cancers.[36] Increasing evidence suggests that adenosine can accumulate at high levels in the tumor microenvironment of major solid tumors,[37,38] where it regulates multiple immune cells through receptor-dependent and/or receptor-independent mechanisms, thereby promoting tumor immune evasion. Similarly, under the complex immune evasion mechanism of EM, ectopic tissues continue to grow under host immune surveillance, and the immune systems of patients with EM generally exhibit significant defects.[39] Therefore, the loss or dysfunction of *ADK* may increase immune suppression by elevating adenosine levels, which may be related to the progression of EM.

EHHADH encodes a bifunctional enzyme that plays a key role in peroxisomal fatty acid β-oxidation, and its primary function is closely related to fatty acid metabolism. In our study, we found that *EHHADH* may act as a protective factor, although no correlation between *EHHADH* and EM has been reported so far. The Ehhadh enzyme encoded by the *EHHADH* gene is a key enzyme in the fatty acid β-oxidation process and is crucial for the normal function of peroxisomes.[40] Existing studies show that mice deficient in *EHHADH* exhibit significant medium-chain 3-hydroxydicarboxylic aciduria when mitochondrial fatty acid oxidation is inhibited. *Ehhadh* knockout mice show increased mRNA and protein expression of enzymes related to cholesterol biosynthesis; however, in female mice, the cholesterol synthesis rate decreases. These results suggest that *EHHADH* plays a critical role in medium-chain dicarboxylic acid metabolism and cholesterol regulation. In addition, *EHHADH* deficiency is closely associated with peroxisomal dysfunction, potentially leading to metabolic abnormalities when large amounts of medium-chain fatty acids are ingested.[41] Such metabolic abnormalities may affect energy production, cholesterol synthesis, and other essential cellular functions. Fatty acids play a significant role in immune cell function, participating in energy provision, biosynthesis, and signal transduction processes.[42] Furthermore, fatty acids are closely related to inflammatory responses.[43] EM is a systemic chronic inflammatory disease, and its occurrence is closely linked to immune responses. Therefore, we hypothesize that *EHHADH* may influence the progression of EM through the regulation of fatty acid metabolism, from an inflammation and immune perspective. However, this mechanism still requires further study and validation.

CNOT7 is one of the essential subunits of the eukaryotic CCR4-NOT protein complex[44] Together with other subunits, it is responsible for the deadenylation of the poly(A) tail of mRNA, thus regulating mRNA degradation.[45] In our study, *CNOT7* was identified as a protective factor, though its specific association with EM has not been clearly defined. Previous studies have shown that CNOT7 binds to the antiproliferative protein B-cell translocation protein 1 (BTG1), which negatively regulates cell proliferation. The antitumor activity of BTG/TOB proteins is mediated by the Caf1a (CNOT7) and Caf1b (CNOT8) deadenylase subunits of the CCR4-NOT complex. The activity of BTG/TOB proteins in mRNA abundance and translation regulation depends on Caf1a/Caf1b, without the need for other CCR4-NOT components.[46] The loss of *CNOT7* may lead to an imbalance in this negative regulation of cell proliferation. Existing studies have shown that in the EM group, the BTG1 mRNA expression levels in both eutopic and ectopic endometrial tissues are significantly lower than those in the control group's eutopic endometrium. Downregulation of BTG1 leads to a significant increase in the migration potential of human endometrial stromal cells.[47] Therefore, we speculate that the loss of *CNOT7* may be associated with the progression of EM.

This study also reveals subtype characteristics and immune infiltration differences associated with different gene expression patterns through unsupervised clustering analysis of endometrial samples. This provides new ideas for addressing the issue

of stratified management in clinical practice and helps optimize personalized treatment strategies based on subtype features.

This study has several notable strengths. First, it integrates GWAS, RNA sequence data, and scRNA-seq data to perform a multidimensional in-depth analysis of genomic variations, gene expression, and cellular aspects. The integration of large-scale and multilayered data greatly enhances the robustness of the analysis results. Second, at the methodological level, a stratified integrated analysis strategy was employed, combining MAGMA, SMR, differential expression analysis, and multiple machine learning techniques. This multidimensional approach comprehensively evaluated the biomarkers, providing independent evidence from various perspectives, such as genomics, causality, expression differences, and predictive models. This created mutually reinforcing multidimensional support, making the role of biomarkers in EM more comprehensive and robust. Finally, the dual validation through ROC curves and expression trends further enhanced the reliability of the model and the diagnostic efficacy of the biomarkers, providing strong support for the precise diagnosis of EM and potential therapeutic targets.

Although this study has many strengths, it also has some limitations. The research mainly relies on data analysis integration, without conducting biological experimental validation. This means that the specific biological functions of these biomarkers in EM cannot be fully confirmed. Future studies could further validate the functions of these biomarkers and their roles in disease mechanisms through cell or animal experiments, thereby providing more direct support for clinical applications.

5. Conclusion

This study successfully identified three core biomarkers for EM: *ADK*, *EHHADH*, and *CNOT7*. Specifically, *ADK*, *EHHADH*, and *CNOT7* may play a protective role. These findings provide significant scientific evidence for the early diagnosis and personalized treatment of EM.

Acknowledgments

The authors express their gratitude to the FinnGen study and the researchers and participants from the datasets provided by the GEO database for supplying invaluable data for this research.

Author contributions

Conceptualization: Ziwei Xie, Yuxin Feng, Yue He, Yingying Lin.
Data curation: Ziwei Xie.
Formal analysis: Ziwei Xie.
Project administration: Ziwei Xie.
Resources: Ziwei Xie.
Software: Ziwei Xie.
Supervision: Ziwei Xie.
Validation: Ziwei Xie.
Visualization: Ziwei Xie.
Writing – original draft: Ziwei Xie, Yuxin Feng, Yue He, Yingying Lin.
Writing – review & editing: Ziwei Xie, Yuxin Feng, Yue He, Yingying Lin.
Funding acquisition: Xiaohong Wang.
Investigation: Xiaohong Wang.
Methodology: Xiaohong Wang.

References

[1] Noh EJ, Kim DJ, Lee JY, et al. Ureaplasma urealyticum infection contributes to the development of pelvic endometriosis through toll-like receptor 2. Front Immunol. 2019;10:2373.
[2] Zondervan KT, Becker CM, Missmer SA. Endometriosis. N Engl J Med. 2020;382:1244–56.
[3] Shafrir AL, Farland LV, Shah DK, et al. Risk for and consequences of endometriosis: a critical epidemiologic review. Best Pract Res Clin Obstet Gynaecol. 2018;51:1–15.
[4] Bulun SE, Yilmaz BD, Sison C, et al. Endometriosis. Endocr Rev. 2019;40:1048–79.
[5] Falcone T, Flyckt R. Clinical management of endometriosis. Obstet Gynecol. 2018;131:557–71.
[6] Ianieri MM, Mautone D, Ceccaroni M. Recurrence in deep infiltrating endometriosis: a systematic review of the literature. J Minim Invasive Gynecol. 2018;25:786–93.
[7] Zhou C, Feng M, Chen Y, et al. Unraveling immunotherapeutic targets for endometriosis: a transcriptomic and single-cell analysis. Front Immunol. 2023;14:1288263.
[8] Kennedy S, Mardon H, Barlow D. Familial endometriosis. J Assist Reprod Genet. 1995;12:32–4.
[9] Painter JN, Nyholt DR, Krause L, et al. Common variants in the CYP2C19 gene are associated with susceptibility to endometriosis. Fertil Steril. 2014;102:496–502.e5.
[10] Audebert A, Lecointre L, Afors K, Koch A, Wattiez A, Akladios C. Adolescent ENDOMETRIOSIS: REPORT OF A SERIES of 55 cases with a focus on clinical presentation and long-term issues. J Minim Invasive Gynecol. 2015;22:834–40.
[11] Treloar SA, O'Connor DT, O'Connor VM, Martin NG. Genetic influences on endometriosis in an Australian twin sample. sueT@qimr.edu.au. Fertil Steril. 1999;71:701–10.
[12] Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42:D1001–1006.
[13] de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015;11:e1004219.
[14] Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48:481–7.
[15] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16:321–32.
[16] Prelaj A, Miskovic V, Zanitti M, et al. Artificial intelligence for predictive biomarker discovery in immuno-oncology: a systematic review. Ann Oncol. 2024;35:29–65.
[17] Addala V, Newell F, Pearson JV, et al. Computational immunogenomic approaches to predict response to cancer immunotherapies. Nat Rev Clin Oncol. 2024;21:28–46.
[18] Kurki MI, Karjalainen J, Palta P, et al; FinnGen. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature. 2023;613:508–18.
[19] Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8:1826.
[20] Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
[21] Xie Z-W, He Y, Feng Y-X, Wang X-H. Identification of programmed cell death-related genes and diagnostic biomarkers in endometriosis using a machine learning and Mendelian randomization approach. Front Endocrinol. 2024;15:1372221.
[22] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7.
[23] GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.
[24] Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenome. eLife. 2018;7:e34408.
[25] Xu J, Yang T, Wu F, Chen T, Wang A, Hou S. A nomogram for predicting prognosis of patients with cervical cerclage. Heliyon. 2023;9:e21147.
[26] Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinf. 2011;12:77.
[27] 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;28:882–3.
[28] Liu J, Li J, Tang Y, et al. Transcriptome analysis combined with Mendelian randomization screening for biomarkers causally associated with diabetic retinopathy. Front Endocrinol (Lausanne). 2024;15:1410066.
[29] Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. Cell. 2019;177:1888–902.e21.

[30] Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16:1289–96.
[31] Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1:417–25.
[32] Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26:1572–3.
[33] Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7.
[34] Park J, Gupta RS. Adenosine kinase and ribokinase – the RK family of proteins. Cell mol life sci. CMLS. 2008;65:2875–96.
[35] Boison D. Adenosine kinase: exploitation for therapeutic gain. Pharmacol Rev. 2013;65:906–43.
[36] Xing J, Zhang J, Wang J. The immune regulatory role of adenosine in the tumor microenvironment. Int J Mol Sci . 2023;24:14928.
[37] Allard B, Longhi MS, Robson SC, Stagg J. The ectonucleotidases CD39 and CD73: novel checkpoint inhibitor targets. Immunol Rev. 2017;276:121–44.
[38] Allard B, Beavis PA, Darcy PK, Stagg J. Immunosuppressive activities of adenosine in cancer. Curr Opin Pharmacol. 2016;29:7–16.
[39] Reis JL, Rosa NN, Martins C, Ângelo-Dias M, Borrego LM, Lima J. The role of NK and T cells in endometriosis. Int J Mol Sci. 2024;25:10141.
[40] Zhao S, Xu W, Jiang W, et al. Regulation of cellular metabolism by protein lysine acetylation. Science (New York, N.Y.). 2010;327:1000–4.
[41] Ranea-Robles P, Violante S, Argmann C, et al. Murine deficiency of peroxisomal L-bifunctional protein (EHHADH) causes medium-chain 3-hydroxydicarboxylic aciduria and perturbs hepatic cholesterol homeostasis. Cell Mol Life Sci. 2021;78:5631–46.
[42] Zhang S, Lv K, Liu Z, Zhao R, Li F. Fatty acid metabolism of immune cells: a new target of tumour immunotherapy. Cell Death Discovery. 2024;10:39.
[43] Phinney SD. Fatty acids, inflammation, and the metabolic syndrome. Am J Clin Nutr. 2005;82:1151–2.
[44] Bartlam M, Yamamoto T. The structural basis for deadenylation by the CCR4-NOT complex. Protein Cell. 2010;1:443–52.
[45] Dai XX, Jiang Y, Gu JH, et al. The CNOT4 subunit of the CCR4-NOT complex is involved in mRNA degradation, efficient DNA damage repair, and XY chromosome crossover during male germ cell meiosis. Adv Sci (Weinheim, Baden-Wurttemberg, Germany). 2021;8:2003636.
[46] Doidge R, Mittal S, Aslam A, Winkler GS. The anti-proliferative activity of BTG/TOB proteins is mediated via the Caf1a (CNOT7) and Caf1b (CNOT8) deadenylase subunits of the Ccr4-not complex. PLoS One. 2012;7:e51331.
[47] Kim JS, Choi YS, Park JH, et al. Role of B-cell translocation gene 1 in the pathogenesis of endometriosis. Int J Mol Sci. 2019;20:3372.