Journal of Assisted Reproduction and Genetics (2023) 40:2439–2451
https://doi.org/10.1007/s10815-023-02903-y

REPRODUCTIVE PHYSIOLOGY AND DISEASE

Identification of potential diagnostic biomarkers and therapeutic targets for endometriosis based on bioinformatics and machine learning analysis

Maryam Hosseini[1] · Behnaz Hammami[1] · Mohammad Kazemi[1,2]

Received: 31 May 2023 / Accepted: 28 July 2023 / Published online: 9 August 2023

Abstract

Purpose Endometriosis (EMs) is a major gynecological condition in women. Due to the absence of definitive symptoms, its early detection is very challenging; thus, it is crucial to find biomarkers to ease its diagnosis and therapy. Here, we aimed to identify potential diagnostic and therapeutic targets for EMs by constructing a regulatory network and using machine learning approaches.

Methods Three Gene Expression Omnibus (GEO) datasets were merged, and differentially expressed genes (DEGS) were identified after preprocessing steps. Using the DEGs, a transcription factor (TF)-mRNA-miRNA regulatory network was constructed, and hub genes were detected based on four different algorithms in CytoHubba. The hub genes were used to build a GaussianNB diagnostic model and also in docking analysis that were performed using Discovery Studio and AutoDock Vina software.

Results A total of 119 DEGs were identified between EMs and non-EMs samples. A regulatory network consisting of 52 mRNAs, 249 miRNAs, and 37 TFs was then constructed. The diagnostic model was introduced using the hub genes selected from the network (*GATA6*, *HMOX1*, *HS3ST1*, *NFASC*, and *PTGIS*) that its area under the curve (AUC) was 0.98 and 0.92 in the training and validation cohorts, respectively. Based on docking analysis, two chemical compounds, rofecoxib and retinoic acid, had potential therapeutic effects on EMs.

Conclusion In conclusion, this study identified potential diagnostic and therapeutic targets for EMs which demand more experimental confirmations.

Keywords Endometriosis · Gene expression · Diagnostic biomarkers · Machine learning · Docking analysis

Introduction

Endometriosis (EMs) is a common gynecological disorder, characterized by the development of endometrial tissue outside the uterus [1, 2]. EMs affects 8% of women of reproductive age, with the main symptoms being pelvic pain, dysmenorrhea, and infertility [3]. Although many hypotheses have been proposed to explain the etiology of EMs, including retrograde menstruation [4], coelomic metaplasia [5], Müllerian remnants [6], and the stem cell theory [7], the pathology of this disorder remains unknown [8, 9]. EMs reduces patients' quality of life and burdens them socially and financially [10]. Due to the lack of conclusive symptoms and non-invasive diagnostic methods, there is a four to 11 years of diagnostic delay for EMs [11]; therefore, it is essential to discover novel biomarkers to facilitate its early diagnosis and individualized treatment [10].

Studies have revealed some changes in the transcriptome profiles of patients with EMs [12]. Microarray data analysis and machine learning techniques have been widely proposed to investigate specific changes in gene expression patterns and pathways in diseases [12, 13]. Gene expression can be regulated by various factors, such as transcription factors (TFs) and microRNAs (miRNAs), and studying all these

✉ Mohammad Kazemi
m_kazemi@med.mui.ac.ir

[1] Department of Genetics and Molecular Biology, School of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran

[2] Reproductive Sciences and Sexual Health Research Center, Isfahan University of Medical Sciences, Isfahan, Iran

components together can help better understand disease causation [14].

The aim of this study was to identify differentially expressed genes (DEGs) in EMs by combining different Gene Expression Omnibus (GEO) microarray datasets. Moreover, our goal was to introduce diagnostic biomarkers for endometriosis by constructing a regulatory network and utilizing machine learning methods. Finally, based on molecular docking analysis, we aimed to identify possible therapeutic compounds for EMs as well.

Materials and methods

Data retrieval, quality control, and normalization

Four gene expression datasets including GSE7305, GSE7307, GSE25628, and GSE11691 were obtained from the NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/). GSE7305 included ten ovarian endometriosis and ten matching normal endometrial tissues. Eight of these samples were in the luteal phase, and two were in the follicular phase of the menstrual cycle. Surgical samples were collected before the use of any drugs, including hormone treatments. GSE7307 included 18 ovarian endometriosis and 23 normal endometrial tissues from women without endometriosis. The menstrual phase of these samples was unknown. GSE25628 contained seven ectopic, nine eutopic, and six control samples collected from participants without endometriosis. All samples were collected during the follicular phase of the menstrual cycle. Collectively, these datasets (GSE7305, GSE7307, and GSE25628) had a total of 35 EMs tissues (eight in the luteal phase, nine in the follicular phase, and 18 samples with an unknown phase) and 48 non-EMs tissues (19 normal endometrial tissues from patients with endometriosis and 29 endometrial tissues from healthy donors, including eight luteal phase, 17 follicular phase, and 23 unknown phase samples). GSE11691 was only used as a validation cohort in machine learning steps (which will be explained in the following sections) that included nine paired samples of peritoneal endometriosis and normal endometrial tissues. Four and five samples were collected during the luteal and follicular phases, respectively. Table 1 lists detailed information about the selected datasets.

Raw CEL files were loaded into R using affy package [15]. Normalized unscaled standard error (NUSE) [16] and relative log expression (RLE) [17] were used to evaluate the quality of the arrays utilizing affyPLM package [18]. Arrays with median NUSE scores less than $1+0.05$ or greater than $1-0.05$ and median RLE scores within -0.1 and 0.1 were considered to have good quality, and those that did not satisfy these requirements were excluded from the analysis. After the preprocessing steps, the remaining raw CEL files were normalized using guanine cytosine robust multi-array analysis (GCRMA) method [19, 20].

Integration of microarray data and batch correction

The GEO datasets (GSE7305, GSE7307, and GSE25628) were combined into one according to their probe IDs, as having a larger dataset can make the statistical analysis more robust and, therefore, the results more reliable [21]. However, different datasets are generated by different groups and under various circumstances, which can result in batch effects [22]. Several techniques, including ComBat empirical Bayes [23] and distance weighted discrimination (DWD) [24] methods, can be used to eliminate batch effects between different datasets. In this study, batch effects were assessed using principal component analysis (PCA) and then adjusted using ComBat function of sva package [25]. After applying batch correction, the data were rechecked for batch effects using PCA, and all subsequent steps were performed using the batch-corrected dataset.

Differential gene expression analysis

Differential gene expression analysis was conducted between EMs and non-EMs tissue samples. DEGs were identified using limma R package [26]. Genes with $|\log_2$ fold change (FC)$|>3$ and adjusted p-values <0.01 were considered as DEGs.

Identification of TFs and miRNAs regulating the DEGs

miRWalk database [27] (http://mirwalk.umm.uni-heidelberg.de/) contains information about both predicted and experimentally validated miRNA-binding sites. This database was

Table 1 Details of the GEO datasets

Datasets	Platform	Submission data	PMID	EMs	Non-EMs	Menstrual cycle phase
GSE7305	GPL570	Mar. 19, 2007	17,640,886	10	10	Eight luteal, two follicular
GSE7307	GPL570	Mar. 19, 2007	–	18	23	Unknown
GSE25628	GPL571	Nov. 28, 2010	23,460,397	7	15	All follicular
GSE11691	GPL96	Jun. 05, 2008	18,688,027	9	9	Four luteal, five follicular

GEO Gene Expression Omnibus, *PMID* PubMed identifier, *EMs* endometriosis

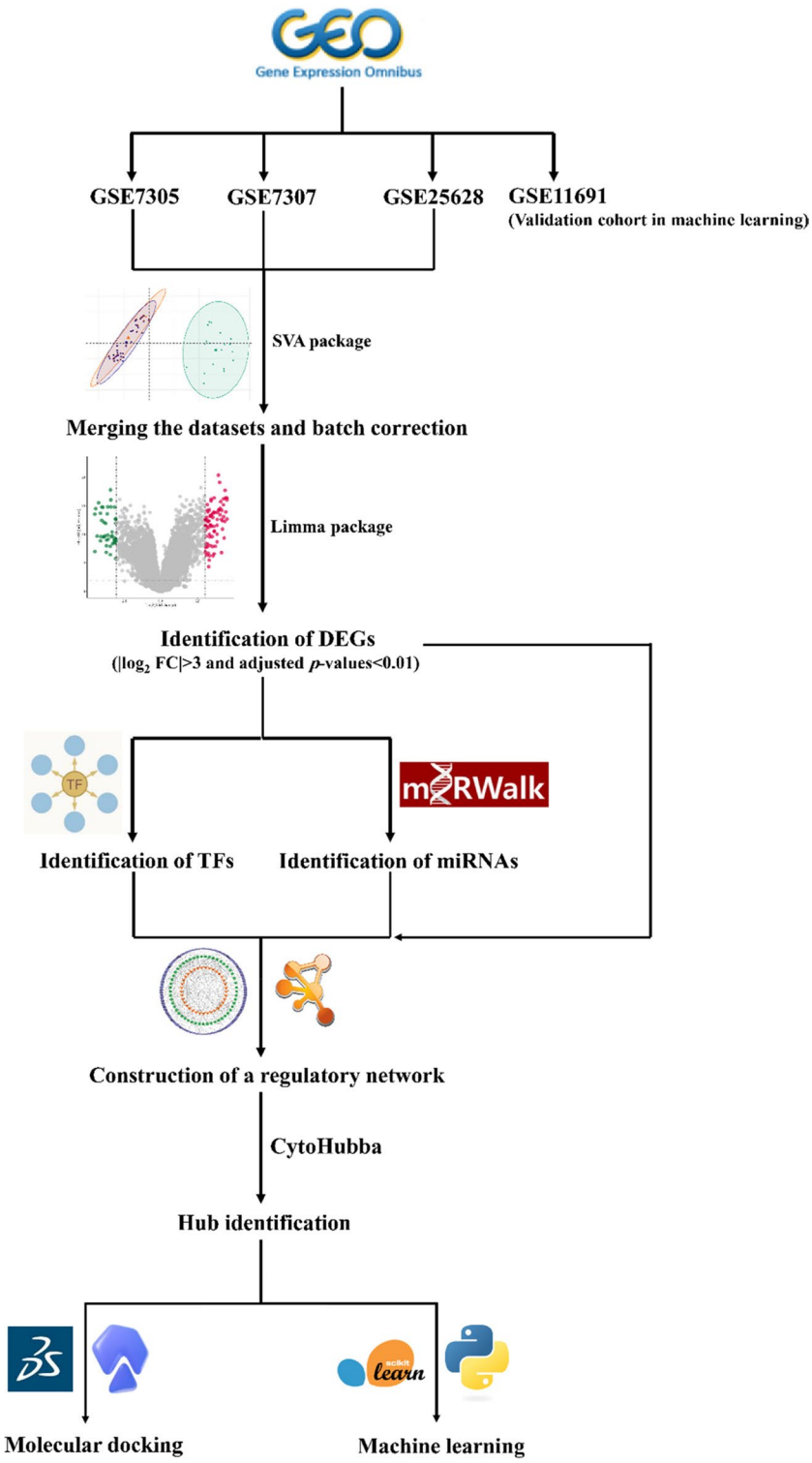


Fig. 1 The workflow for identification of EMs diagnostic and therapeutic targets. Three GEO datasets were merged, and batch correction was conducted. DEGs were identified and then used to predict potential TFs and miRNAs that regulate them. Hub genes selected from the TF-mRNA-miRNA regulatory network were used for molecular docking and machine learning analyses. EMs, endometriosis; GEO, Gene Expression Omnibus; DEG, differentially expressed gene; TF, transcription factor; miRNA, microRNA; FC, fold change

utilized to identify miRNAs that bind to the 3′ untranslated region (3′UTR) of DEGs. Only those interactions that were experimentally verified by miRTarBase database [28] and had a binding score > 0.95 were retained. The potential TFs that regulate the DEGs were also predicted using transcriptional regulatory relationships unraveled by sentence-based text mining (TRRUST) database [29] (https://www.grnpedia.org/trrust/).

TF-mRNA-miRNA network construction and finding hub genes

After predicting mRNA-miRNA and TF-mRNA interactions, a TF-mRNA-miRNA regulatory network was constructed and visualized using Cytoscape software (version 3.9.1; http://www.cytoscape.org/). CytoHubba is a Cytoscape plugin that can identify the core genes of a protein–protein interaction (PPI) network based on 11 different algorithms [30]. In the current study, four of these 11 algorithms, including degree, betweenness, closeness, and maximal clique centrality (MCC), were used to identify the top 10 core mRNAs, and those which were common among them were defined as hub genes.

Building a diagnostic model for EMs

Hub genes were used to build a diagnostic model for EMs to discriminate between EMs and non-EMs samples. Using the combined dataset resulting from merging GSE7305, GSE7307, and GSE25628 datasets as the training cohort and GaussianNB algorithm, a diagnostic model was constructed. The performance of the model was evaluated on the training cohort and an external validation cohort (GSE30601) using fivefold cross-validation with the cross_val_score function in Python. In fivefold cross-validation, data is randomly divided into five partitions. Each time, four partitions will be used for training the model and one for checking its performance. For cross-validation, we used Stratified K-Fold method (StratifiedKFold function in Python), which preserves the proportion of labels in each fold the same as the original dataset. AUC, F1-score, precision, and recall were the performance metrics that were calculated to assess the model's performance. Python packages scikit-learn [31] and matplotlib [32] were used to perform machine learning analyses and visualizations.

Identification of drug candidates and prediction of drug-like properties

DSigDB database provides a direct link between genes and drugs for drug development and translational studies [33] which is accessible through Enrichr web server (https://maayanlab.cloud/Enrichr/). Hub genes were uploaded to Enrichr, and top five chemical compounds that targeted them were selected based on adjusted p-values < 0.01. Drug-like properties of the selected compounds were evaluated using SwissADME database (http://www.swissadme.ch/), and those that did not satisfy a defined drug-like property were excluded from the analysis. Based on the laws of Lipinski [34], Egan [35, 36], and Veber [37], selection criteria were as follows: molecular weight (MW) < 500, number of hydrogen (H)-bond acceptors (nOHNH) ≤ 10, number of H-bond donors (nON) ≤ 5, water partition coefficient (WLOGP) ≤ 5.88, topological surface area (TPSA) < 140, and number of rotatable bonds (nrotb) ≤ 10. High gastrointestinal (GI) absorption, not being a P-glycoprotein (P-gp) substrate, and CYP2D6 or CYP3A4 inhibitor were also considered as other filters for selecting the compounds.

Molecular docking

To investigate the interactions between hub genes and selected chemical compounds, crystal structures of the target proteins were retrieved from the research collaboratory for structural bioinformatic protein data bank (RCSB PDB, http://www.rcsb.org/) [38], in PDB format with selection criteria of resolution < 2.5 Å. In addition, from PubChem database (https://pubchem.ncbi.nlm.nih.gov/), the 3D structures of the selected chemical compounds were downloaded in structure data file (SDF) format and then converted to PDB format using Discovery Studio software [39]. The target proteins were prepared for docking by removing water molecules, adding polar hydrogen atoms and Kollman charges to their structures [40, 41]. The active sites of the target proteins were predicted using Discovery Studio software. Utilizing AutoDock Vina, the receptor and ligand structures were converted into PDBQT format, which is required for docking. Finally, docking was carried out using AutoDock Vina as previously mentioned [42–44], and Discovery Studio was used to visualize the docking results.

Statistical analysis

R software 4.1.3 was used to perform the calculations. Python 3.9.7 on Anaconda 4.10.3 was used to implement

Table 2 Quality assessment of the datasets using NUSE and RLE plots

Datasets	NUSE > 1.05	NUSE < 0.95	RLE > 0.1	RLE < −0.1
GSE7305	–	–		–
GSE7307	GSM176127, GSM176236, GSM176238, GSM176240	–	GSM176240	–
GSE25628	GSM629719, GSM629733	–	GSM629723, GSM629733	–

NUSE normalized unscaled standard error, *RLE* relative log expression

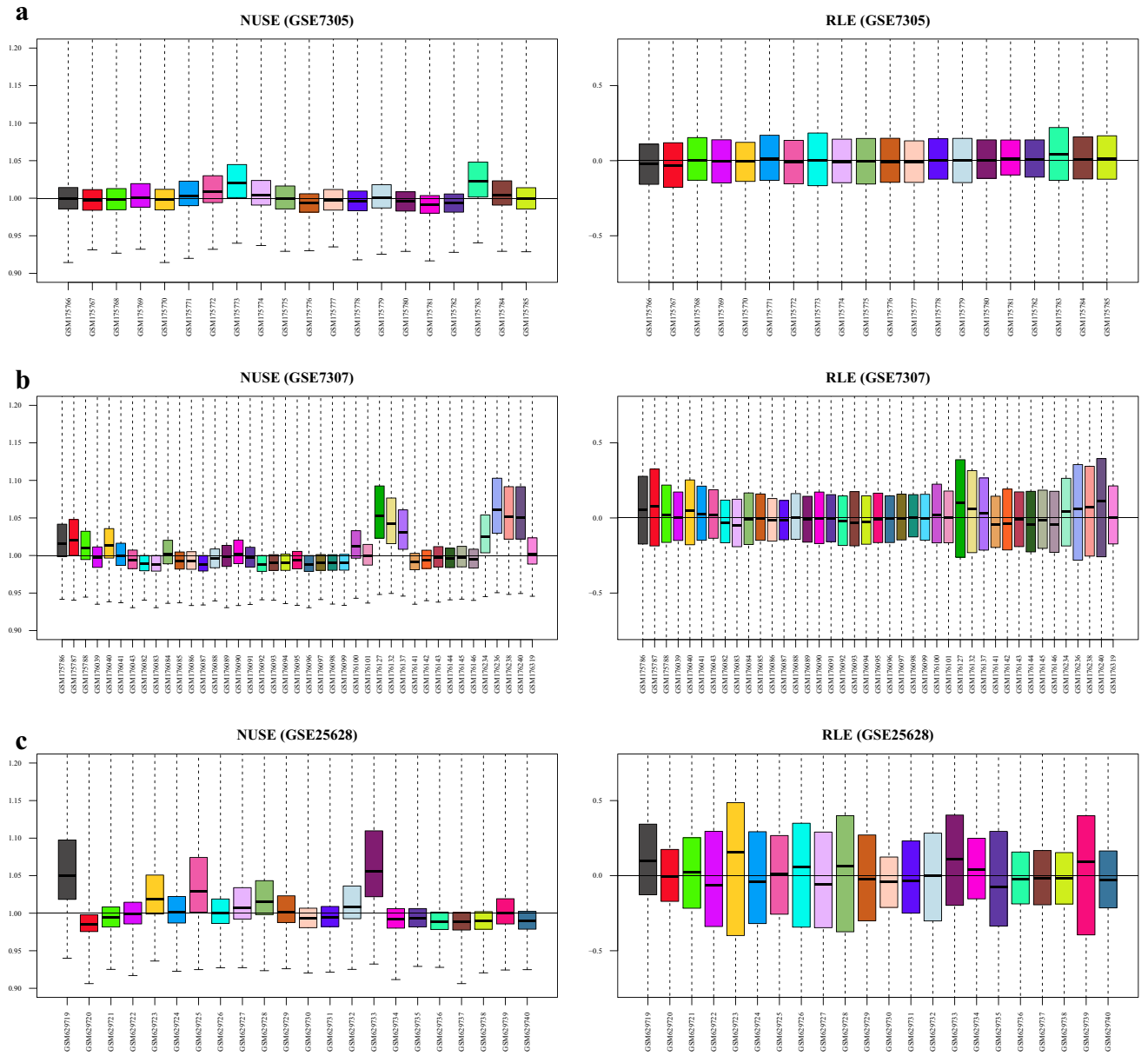


Fig. 2 Quality control of the gene expression datasets. The quality of the samples in each dataset was assessed using NUSE and RLE plots. Those samples retained for the analyses which had 0.95 < median NUSE < 1.05 and − 0.1 < median RLE < 0.1. Thus, zero, four, and three samples were deleted from GSE7305 (**a**), GSE7307 (**b**), and GSE25628 (**c**), respectively. NUSE, normalized unscaled standard error; RLE, relative log expression

the machine learning sections. With the Benjamini/Hochberg method, all *p*-values were corrected.

Results

Quality assessment and normalization of the data

The workflow of this study is schematically shown in Fig. 1. First, the quality of the datasets was evaluated using NUSE and RLE plots, and their value distributions were visualized using boxplots (Fig. 2a–c). The NUSE and RLE values should be relatively close to one and zero, respectively. In the current study, samples with 0.95 < median NUSE < 1.05 and − 0.1 < median RLE < 0.1 were considered to have good quality. Based on these cutoffs, four and three samples were unqualified in GSE7307 and GSE25628, respectively, and were removed for subsequent steps; however, all samples in GSE7305 dataset met the aforementioned criteria (Table 2). After quality assessment of the datasets and removal of

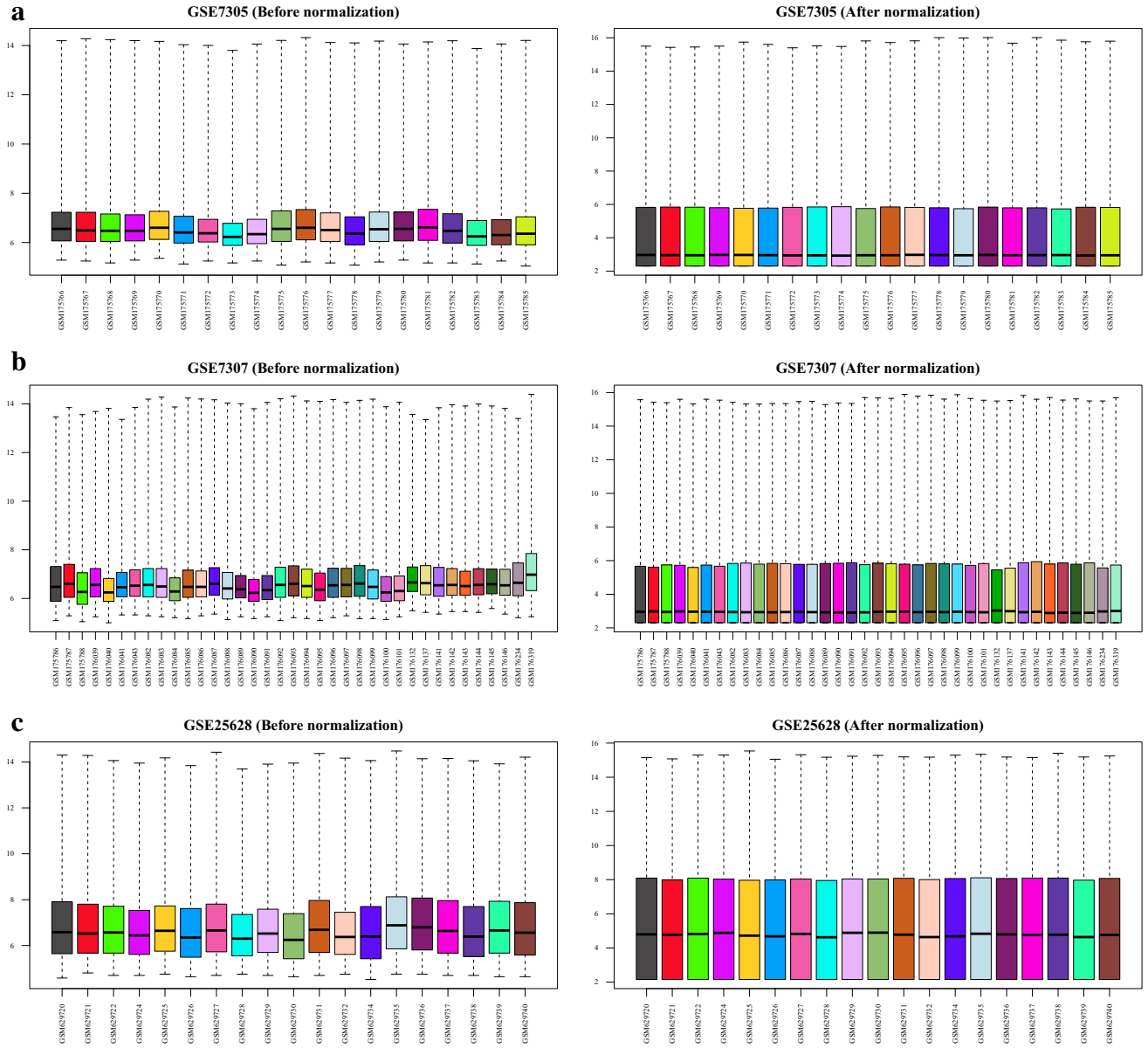


Fig. 3 Normalization of the gene expression datasets. GCRMA method was used to normalize the gene expression datasets. The left and right boxplots show the statues of GSE7305 (**a**), GSE7307 (**b**), and GSE25628 (**c**) datasets before and after normalization. GCRMA, guanine cytosine robust multi-array analysis

unqualified samples, each dataset was normalized using GCRMA method. Figure 3 shows boxplots of the unnormalized and GCRMA-normalized expression values.

Batch removal and identification of DEGs

Three datasets (GSE7305, GSE7307, and GSE25628) were combined, and PCA was conducted to check for batch effects (Fig. 4a). Since the datasets were not completely continuous, in particular, GSE25628 was significantly separated from the others, so batch effect removal was performed using ComBat function. PCA analysis was carried out again on the corrected data, and as shown in Fig. 4b, following the correction, there was no longer a separation between the three datasets. This combined dataset was used in all subsequent analyses and as the training cohort in machine learning steps.

After batch correction, a differential expression analysis was performed on the data. One hundred nineteen genes (78 upregulated and 41 downregulated) were considered DEGs since they met the criteria of $\log_2 FC > 3$ along with adjusted p-values < 0.01 (Fig. 4c). Hierarchical

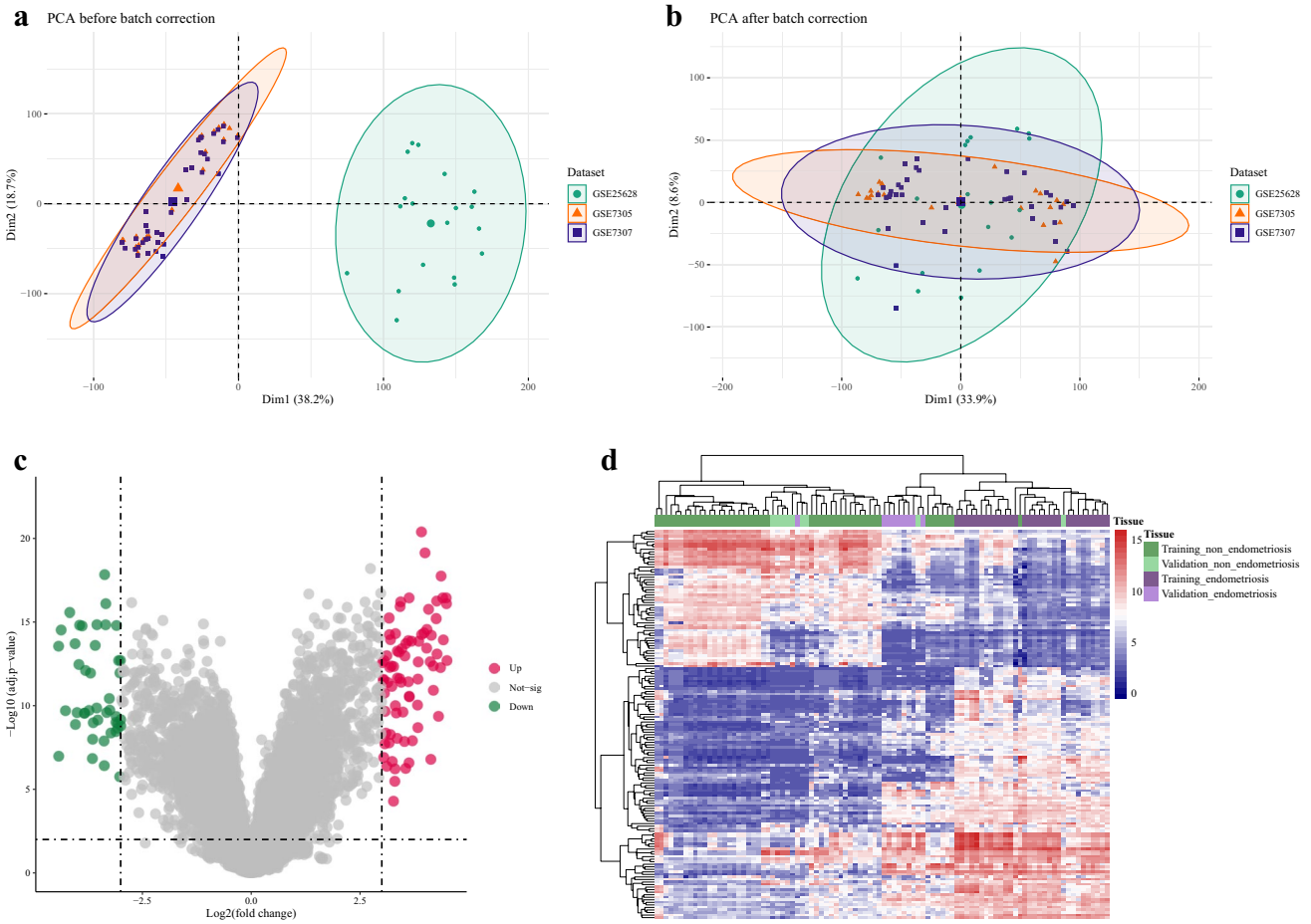


Fig. 4 Correcting the batch effect and identification of DEGs. The PCA plot of the merged GEO datasets (GSE7305 (orange, triangle), GSE7307 (blue, square), and GSE25628 (green, circle)), before batch correction. GSE25628 is separated from the others (**a**). The PCA plot of the merged GEO datasets (GSE7305 (orange, triangle), GSE7307 (blue, square), and GSE25628 (green, circle)), after batch correction. The three datasets are no longer separated from one another (**b**). Volcano plot for DEGs. DEGs were found with the criteria of $|\log_2 FC| > 3$ and adjusted p-values < 0.01. Red and green dots represent upregulated and downregulated DEGs, respectively, and the gray ones were not significant according to the defined criteria (**c**). Hierarchical clustering of 119 DEGs in 44 EMs and 39 non-EMs samples from the merged datasets and 9 EMs and 9 non-EMs samples from GSE11691. EMs and non-EMs samples were clustered together based on these DEGs (**d**). DEG, differentially expressed gene; PCA, principal component analysis; FC, fold change; EMs, endometriosis

Table 3 Finding hub genes based on four different algorithms of CytoHubba

MCC	Degree	Closeness	Betweenness	Common mRNAs
NFASC	*ESR1*	*ESR1*	*ESR1*	
PTGIS	*NFASC*	*TYMS*	*GATA6*	*GATA6*
HS3ST1	*PTGIS*	*HMOX1*	*HMOX1*	
HMOX1	*HS3ST1*	*SP1*	*PTGIS*	*HMOX1*
GATA6	*HMOX1*	*HS3ST1*	*HS3ST1*	
KLF2	*GATA6*	*TP53*	*NFASC*	*HS3ST1*
KLHDC8A	*KLF2*	*PTGIS*	*PRELP*	*NFASC*
PRELP	*KLHDC8A*	*GATA6*	*RRM2*	
RPM2	*PRELP*	*NFASC*	*SP1*	*PTGIS*
VCAM1	*RRM2*	*KLF2*	*TP53*	

MCC maximal clique centrality

clustering using these DEGs and both training and validation (GSE11691) cohorts showed that EMs and non-EMs samples could be separated from each other based on these DEGs (Fig. 4d).

Construction of a TF-mRNA-miRNA regulatory network and hub gene screening

To investigate the regulatory mechanisms in EMs, at first, potential miRNAs and TFs that target the DEGs were predicted (Online Resources 1 & 2). Here, a total of 324 miRNA-mRNA and 109 TF-mRNA pairs were predicted and then combined to form a TF-mRNA-miRNA regulatory network (Fig. 5). A total of 394 interactions were shaped in this network, involving 52 mRNAs, 249 miRNAs, and 37 TFs. The top ten mRNAs were obtained using CytoHubba

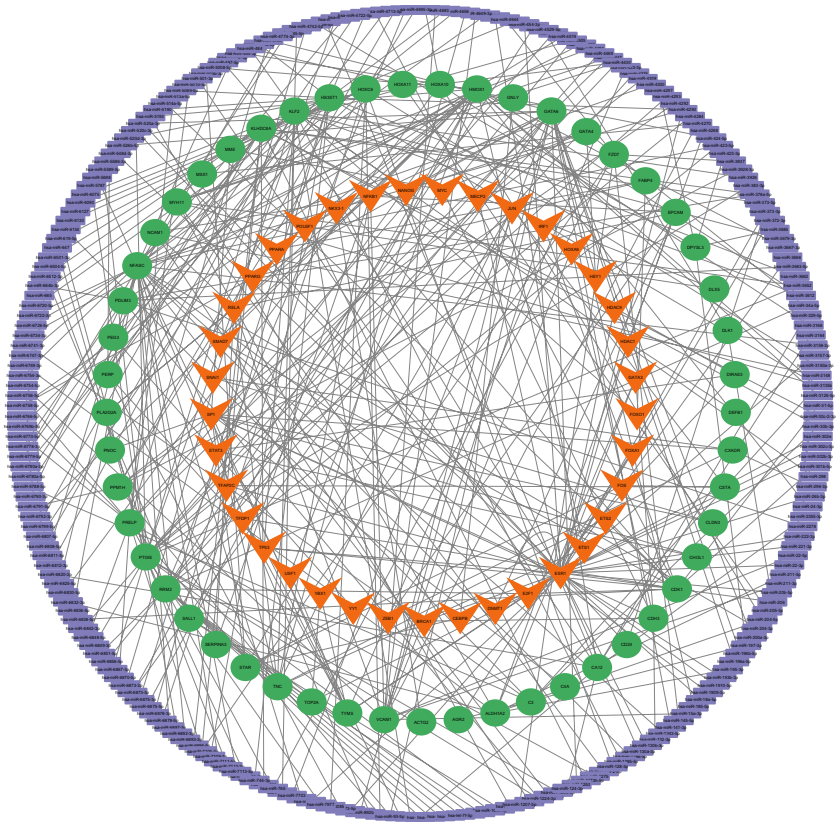


Fig. 5 Construction of a regulatory network for EMs. Regulatory network including mRNAs (green), miRNAs (purple), and TFs (orange). EMs, endometriosis; miRNA, microRNA; TF, transcription factor

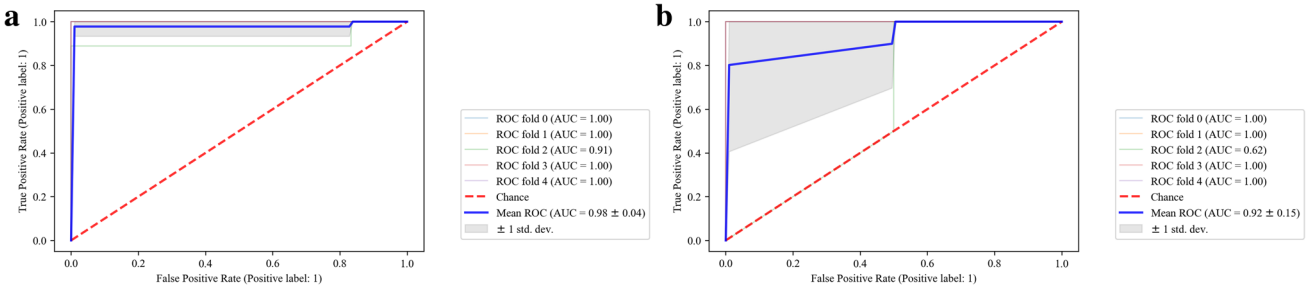


Fig. 6 Model construction and evaluation. Evaluation of the GaussianNB model's performance on training set with fivefold cross-validation. AUC of the model was 0.98 (**a**). Evaluation of the GaussianNB model's performance on the validation cohort with fivefold cross-validation. AUC of the model was 0.92 (**b**). AUC, area under the curve; ROC, receiver operating characteristic

Table 4 Performance of the diagnostic model on the training and validation cohorts

Metric	Training cohort with fivefold cross-validation	GSE30601 with fivefold cross-validation
AUC	0.98	0.92
F1-score	0.94	0.89
Precision	0.98	0.93
Recall	0.91	0.90

AUC area under the curve

plugin, based on four different algorithms: degree, closeness, betweenness, and MCC. The intersecting mRNAs of these four algorithms were considered hub genes and used in the subsequent steps (Table 3).

Building a diagnostic model for EMs

For model construction, a GaussianNB diagnostic model was established using the merged dataset as the training cohort. Model performance was evaluated on the training cohort with fivefold cross-validation. Based on this approach, the AUC of the model was 0.98 for the training

cohort (Fig. 6a). To investigate the reproducibility of the results, the model's performance was further checked using an external validation cohort (GSE30601) with fivefold cross-validation (Fig. 6b). Results of this analysis were consistent with those of the training cohort, showing that these hub genes and the constructed diagnostic model can discriminate between EMs and non-EMs samples. As shown in Table 4, to assess the efficiency of the model more precisely, in addition to AUC, other performance metrics, including F1-score, precision, and recall, were also calculated (Table 4).

Retrieving candidate chemical compounds

The five hub genes were uploaded to Enrichr, which provides a list of potential molecules that target genes based on data from DSigDB database. The top five chemical compounds were selected based on adjusted p-values < 0.01. The selected compounds were ns-398, epoprostenol, rofecoxib, retinoic acid, and acrolein (Table 5).

Molecular docking analysis

First, crystal structures of the five hub genes (*GATA6*, *HMOX1*, *HS3ST1*, *NFASC*, and *PTGIS*) were retrieved from PDB database. The crystal structure of *GATA6* was not available in PDB database, and for *NFASC*, there was no structure with resolution < 2.5 Å; thus, molecular docking was only performed for *HMOX1* (6EHA), *HS3ST1* (1ZRH), and *PTGIS* (3B6H). Next, the simplified molecular input line entry specification (SMILES) IDs of the chemical compounds were used as inputs to SwissADME database to evaluate the drug-like of the selected compounds. Based on the criteria mentioned in the methods, only two of these compounds, rofecoxib and retinoic acid, were suitable for docking (Online Resource 3); therefore, their SDF structures were downloaded from PubChem database. Molecular docking was performed to evaluate the binding affinities of the two selected chemical compounds to the three hub targets. A lower affinity score indicates a stronger binding ability, and in this study, a binding energy < − 5 kJ/mol was used as the screening criterion. These findings indicate that the compounds and target proteins interact with each other with a binding energy of less than − 5 kJ/mol (Table 6); therefore, they have low conformational energy, a stable structure, and high binding activity. Based on these results, different interactions such as hydrogen and van der Waals bonds were formed between the compounds and amino acid residues (Fig. 7).

Table 6 Results of docking analysis and the binding energies between drug and target

Protein	Chemical compounds	Binding energy (kcal/mol)
HMOX1	Rofecoxib	− 8.4
HMOX1	Retinoic acid	− 6.7
HS3ST1	Rofecoxib	− 10
HS3ST1	Retinoic acid	− 7.5
PTGIS	Rofecoxib	− 8.1
PTGIS	Retinoic acid	− 7.6

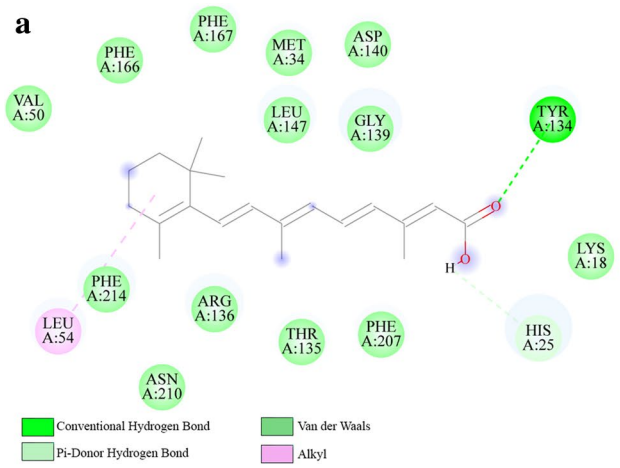
Discussion

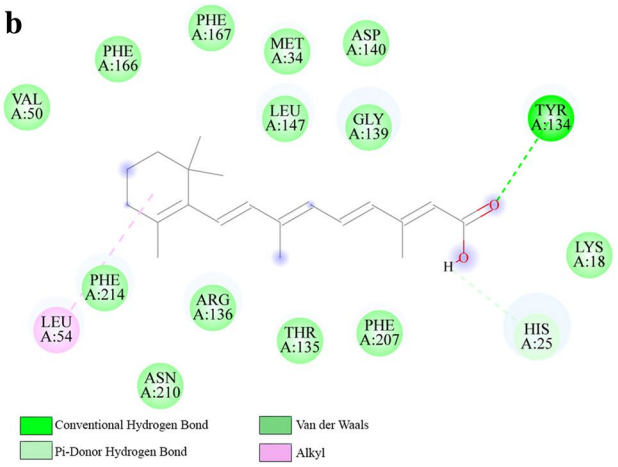
A total of 190 million women suffer from EMs worldwide [45]. It affects women's quality of life in different aspects such as the chance of education or finding a stable job [46]. EMs has recently been linked to an increased risk of various conditions such as cancer [47] and cardiovascular [48] diseases. Early diagnosis and treatment of this disease is very challenging because the pathology underlying its development is still unknown [49]. Although a wide range of biomarkers have been introduced for the early detection of EMs, there is still a significant gap in identifying sensitive and specific biomarkers for this condition [50]. Recently, the biological process of EMs was thought to be significantly influenced by miRNAs, and TFs are also believed to be strongly linked to the onset of the illness, but less research has been done on the regulatory network of these molecules in EMs [51].

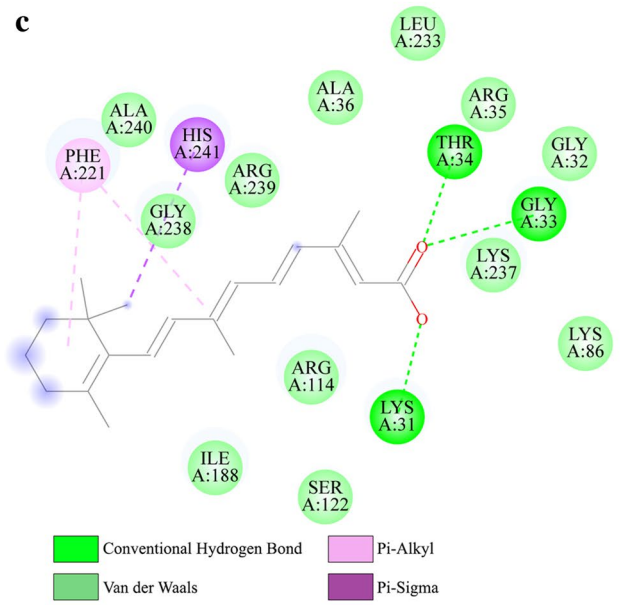
In the present study, expression data for EMs were retrieved from GEO database. Three datasets were combined

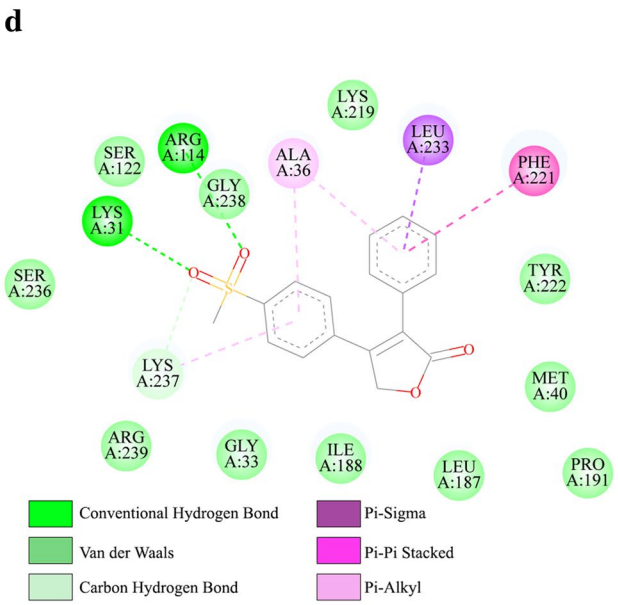
Table 5 Identification of chemical compounds (top five) based on gene-drug interactions

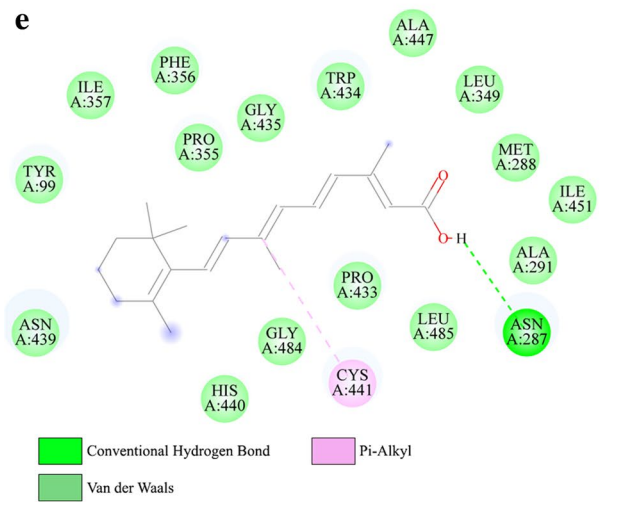
Chemical compounds	Adjusted p-value	Combined score	Common mRNAs
Ns-398	0.040163215	1644.950081	*GATA6*, *HMOX1*
Epoprostenol	0.040163215	1590.730283	*PTGIS*, *HMOX1*
Rofecoxib	0.040163215	1139.724512	*NFASC*, *PTGIS*
Retinoic acid	0.040163215	608,943.1725	*NFASC*, *PTGIS*, *GATA6*, *HMOX1*, *HS3ST1*
ACROLEIN	0.040163215	739.8545889	*PTGIS*, *HMOX1*

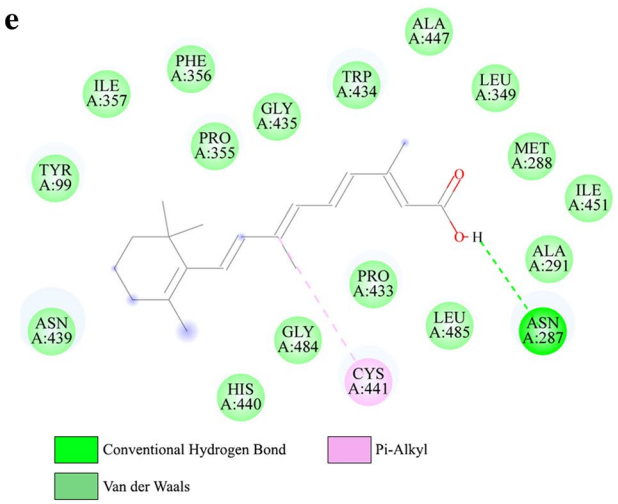

a
PHE A:167
PHE A:166
MET A:34
ASP A:140
VAL A:50
LEU A:147
GLY A:139
TYR A:134
LYS A:18
PHE A:214
ARG A:136
LEU A:54
THR A:135
PHE A:207
HIS A:25
ASN A:210
Conventional Hydrogen Bond
Van der Waals
Pi-Donor Hydrogen Bond
Alkyl


b
PHE A:167
PHE A:166
MET A:34
ASP A:140
VAL A:50
LEU A:147
GLY A:139
TYR A:134
LYS A:18
PHE A:214
ARG A:136
LEU A:54
THR A:135
PHE A:207
HIS A:25
ASN A:210
Conventional Hydrogen Bond
Van der Waals
Pi-Donor Hydrogen Bond
Alkyl


c
LEU A:233
ALA A:36
ARG A:35
ALA A:240
HIS A:241
THR A:34
GLY A:32
PHE A:221
ARG A:239
GLY A:33
GLY A:238
LYS A:237
LYS A:86
ARG A:114
LYS A:31
ILE A:188
SER A:122
Conventional Hydrogen Bond
Pi-Alkyl
Van der Waals
Pi-Sigma


d
LYS A:219
LEU A:233
SER A:122
ARG A:114
ALA A:36
PHE A:221
GLY A:238
LYS A:31
SER A:236
TYR A:222
LYS A:237
MET A:40
ARG A:239
GLY A:33
ILE A:188
LEU A:187
PRO A:191
Conventional Hydrogen Bond
Pi-Sigma
Van der Waals
Pi-Pi Stacked
Carbon Hydrogen Bond
Pi-Alkyl


e
ALA A:447
PHE A:356
ILE A:357
TRP A:434
LEU A:349
GLY A:435
PRO A:355
MET A:288
TYR A:99
ILE A:451
ALA A:291
PRO A:433
ASN A:439
LEU A:485
GLY A:484
ASN A:287
CYS A:441
HIS A:440
Conventional Hydrogen Bond
Pi-Alkyl
Van der Waals


e
ALA A:447
PHE A:356
ILE A:357
TRP A:434
LEU A:349
GLY A:435
PRO A:355
MET A:288
TYR A:99
ILE A:451
ALA A:291
PRO A:433
ASN A:439
LEU A:485
GLY A:484
ASN A:287
CYS A:441
HIS A:440
Conventional Hydrogen Bond
Pi-Alkyl
Van der Waals

◂**Fig. 7** Molecular docking results of the interactions between ligand and specific amino acid residues in the active site of the protein. The molecular docking of HMOX1-retinoic acid (**a**). The molecular docking of HMOX1-rofecoxib (**b**). The molecular docking of HS3ST1-retinoic acid (**c**). The molecular docking of HS3ST1-rofecoxib (**d**). The molecular docking of PTGIS-retinoic acid (**e**). The molecular docking of PTGIS-rofecoxib (**f**)

to boost the sample size and accuracy of the results. DEGs between EMs and non-EMs samples were identified, and a TF-mRNA-miRNA regulatory network was constructed. Network analysis showed that miR-200a-3p, miR-196b-5p, and miR-141-3p targeted *GATA6*. According to previous studies, EMs patients have reduced levels of mir-200a and mir-141, which increase the epithelial to mesenchymal transition (EMT) process, invasion, and motility of endometrial cells [52]. Several TFs were predicted to be associated with the pathogenesis of EMs. Based on the network, *STAT3*, *NFKB1*, and *RELA* targeted *HMOX1*, and *NANOG* targeted *GATA6*. Kim et al. found *STAT3* abnormal activation in the eutopic endometrium of EMs patients [53]. Nuclear factor-kB (NF-kB) regulates cell proliferation and angiogenesis in a variety of cell types that are involved in the development of EMs [54]. According to Song et al., *NANOG* is overexpressed in women of reproductive age with ovarian EMs [55].

Five hub mRNAs, selected from the regulatory network based on different criteria, were used to build a GaussianNB diagnostic model for EMs. Different metrics were then calculated to assess the model performance on the training and validation sets with fivefold cross-validation. Based on this approach, the AUC of the model on the training and validation datasets was 0.98 and 0.92, respectively, demonstrating its excellent ability to differentiate between samples with and without EMs.

The five hub genes were *GATA6*, *HMOX1*, *HS3ST1*, *NFASC*, and *PTGIS*. Some of these genes have previously been reported to be involved in the pathogenesis of EMs. *GATA6* is one of the necessary components for converting endometrial stromal cells into cells that behave similarly to endometriotic stromal cells [56]. In a study conducted by Izawa et al., *GATA6* was introduced as a diagnostic marker for EMs and its body sequence may behave as an active enhancer under the influence of DNA methylation [57]. *HMOX1* has a significant impact on the etiopathogenesis of EMs, potentially by promoting endometriotic cell survival in ectopic locations [58]. *NFASC* has been demonstrated to cause neuropathic hyperalgesia, which worsens pelvic pain in EMs patients and may serve as a new biomarker and therapeutic target for EMs [59]. According to Bae et al., *PTGIS* expression is upregulated in endometrial lesions. This gene is involved in signaling pathways such as TLR4/NF-kB, Wnt/frizzled, and estrogen receptors [60].

Molecular docking revealed that hub proteins HMOX1, HS3ST1, and PTGIS directly interacted with rofecoxib and retinoic acid. Several studies have shown that rofecoxib causes atrophy and regression of the endometriotic lesions [61, 62]. Furthermore, long-term rofecoxib therapy may decrease the chronic pelvic pain associated with EMs [63]. However, rofecoxib had to be removed from the market due to serious adverse effects like myocardial infarction and stroke [64]. Retinoic acid production within the endometrial tissue of the uterine is essential for normal endometrial cell differentiation, activity, and decidualization [65, 66]. Abnormal retinoic acid metabolism can result in the development of EMs lesions [65]. The cell surface receptor stimulated by retinoic acid 6 (STRA6), which is the major receptor for retinol absorption, is downregulated in endometriotic stromal cells compared to normal endometrial cells [67]. These findings imply that rofecoxib and retinoic acid could be potential therapeutic options for EMs.

It should be noted that the current study has a number of limitations. First, the sample size was small, even after combining the three GEO datasets, and this limitation may affect the interpretability of the results. Second, these biomarkers were detected based on tissue samples, and in future studies, it will be necessary to evaluate their effectiveness in blood samples to ensure that these genes can aid in the non-invasive identification of endometriosis. Third, the samples used in this study were collected at different phases of the menstrual cycle due to the lack of datasets with proper characteristics. Fourth, further in vivo and in vitro evaluations (with larger sample sizes) are required to validate the results of this study.

In summary, using bioinformatics and machine learning approaches, a set of five genes were identified that have the potential to help with the early detection of EMs. Based on different metrics, like AUC, these biomarkers have high sensitivity and specificity. Small drug molecules associated with these hub genes have also been identified. Further experimental evaluations and clinical validations are required to validate these results.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s10815-023-02903-y.

Author contribution Maryam Hosseini: methodology, software, formal analysis, investigation, data curation, writing—original draft, writing—review and editing, visualization Behnaz Hammami: methodology, formal analysis, writing—original draft, writing—review and editing. Mohammad Kazemi: conceptualization, methodology, investigation, resources, writing—original draft, writing—review and editing, validation, project administration.

Data availability The GEO datasets used in this study are available at GEO (https://www.ncbi.nlm.nih.gov/geo/) database. The list of the datasets is provided in Table 1.

Declarations

Competing interests The authors declare no competing interests.

References

1. Dai F-F, Bao A-Y, Luo B, Zeng Z-H, Pu X-L, Wang Y-Q, et al. Identification of differentially expressed genes and signaling pathways involved in endometriosis by integrated bioinformatics analysis. Exp Ther Med. 2019;19:264 (Spandidos Publications).
2. Zhang Z, Ruan L, Lu M, Yao X. Analysis of key candidate genes and pathways of endometriosis pathophysiology by a genomics-bioinformatics approach. Gynecol Endocrinol. 2019;35:576–81.
3. Ye Z, Meng Q, Zhang W, He J, Zhao H, Yu C, et al. Exploration of the shared gene and molecular mechanisms between endometriosis and recurrent pregnancy loss. Front Vet Sci. 2022;9:867405.
4. Mathew AG. A case exemplifying Sampson's theory of the aetiology of endometriosis. Aust N Z J Obstet Gynaecol. 1963;3:159–61.
5. Matsuura K, Ohtake H, Katabuchi H, Okamura H. Coelomic metaplasia theory of endometriosis: evidence from in vivo studies and an in vitro experimental model. Gynecol Obstet Invest. 1999;47(Suppl 1):18–22.
6. Ugur M, Turan C, Mungan T, Kuscu E, Senoz S, Agis HT, et al. Endometriosis in association with Mullerian anomalies. Gynecol Obstet Invest. 1995;40:261–4.
7. Maruyama T. A revised stem cell theory for the pathogenesis of endometriosis. J Pers Med. 2022;12(2):216 (Multidisciplinary Digital Publishing Institute (MDPI)).
8. Bai J, Wang B, Wang T, Ren W. Identification of functional lncRNAs associated with ovarian endometriosis based on a ceRNA network. Front Genet. 2021;12:534054 (Frontiers Media S.A.).
9. Wu J, Fang X, Xia X. Identification of key genes and pathways associated with endometriosis by weighted gene co-expression network analysis. Int J Med Sci. 2021;18:3425–36 (Ivyspring International Publisher).
10 Cui D, Liu Y, Ma J, Lin K, Xu K, Lin J. Identification of key genes and pathways in endometriosis by integrated expression profiles analysis. PeerJ. 2020;8:e10171 (PeerJ Inc.).
11. Akter S, Xu D, Nagel SC, Bromfield JJ, Pelch K, Wilshire GB, et al. Machine learning classifiers for endometriosis using transcriptomics and methylomics data. Front Genet. 2019;10:766 (Frontiers Media S.A.).
12. Akter S, Xu D, Nagel SC, Bromfield JJ, Pelch KE, Wilshire GB, et al. GenomeForest: an ensemble machine learning classifier for endometriosis. AMIA Jt Summits Transl Sci. 2020;2020:33–42 (American Medical Informatics Association).
13 Miao C, Chen Y, Fang X, Zhao Y, Wang R, Zhang Q. Identification of the shared gene signatures and pathways between polycystic ovary syndrome and endometrial cancer: an omics data based combined approach. PLoS One. 2022;17:e0271380.
14. Zhang HM, Kuang S, Xiong X, Gao T, Liu C, Guo AY. Transcription factor and microRNA co-regulatory loops: important regulatory motifs in biological processes and diseases. Brief Bioinform. 2013;16:45–58 (Oxford Academic).
15 Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20:307–15 (Oxford Academic).
16. Brettschneider J, Bolstad B, Collin F, Speed T. Quality assessment for short oligonucleotide microarray data. Technometrics. 2008;50:241–64 (Taylor & Francis).
17. Heber S, Sick B. Quality assessment of Affymetrix GeneChip data. OMICS. 2006;10:358–68.
18 Bolstad BM, Collin F, Brettschneider J, Simpson K, Cope L, Irizarry RA, et al. Quality assessment of Affymetrix GeneChip data. Bioinforma Comput Biol Solut Using R Bioconductor. New York, NY: Springer; 2005. p. 33–47.
19. Harr B, Schlötterer C. Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons. Nucleic Acids Res. 2006;34:1–8 (Oxford Academic).
20. Bioconductor - gcrma [Internet]. Available from: https://www.bioconductor.org/packages/release/bioc/html/gcrma.html. Accessed 21 Feb 2023
21. Ai D, Wang Y, Li X, Pan H. Colorectal cancer prediction based on weighted gene co-expression network analysis and variational auto-encoder. Biomolecules. 2020;10:1–11 (Multidisciplinary Digital Publishing Institute).
22. Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, et al. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenomics. 2010;10:278–91 (J Nature Publishing Group).
23. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27 (Oxford Academic).
24. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, et al. Adjustment of systematic microarray data biases. Bioinformatics. 2004;20:105–14 (Oxford Academic).
25. 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.
26. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:47 (Oxford Academic).
27 Dweep H, Gretz N. MiRWalk2.0: a comprehensive atlas of microRNA-target interactions. Nat Methods. 2015;12(8):697–697 (Nature Publishing Group).
28. Huang HY, Lin YCD, Cui S, Huang Y, Tang Y, Xu J, et al. MiRTarBase update 2022: an informative resource for experimentally validated miRNA-target interactions. Nucleic Acids Res. 2022;50:D222–30 (Oxford Academic).
29. Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018;46:D380–6 (Oxford Academic).
30 Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol. 2014;8:S11 (BioMed Central).
31. Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30 (Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot).
32. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007;9:90–5 (IEEE Computer Society).
33. Yoo M, Shin J, Kim J, Ryall KA, Lee K, Lee S, et al. DSigDB: drug signatures database for gene set analysis. Bioinformatics. 2015;31:3069–71 (Oxford University Press).
34. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001;46:3–26 (Elsevier).
35. Egan WJ, Merz KM, Baldwin JJ. Prediction of drug absorption using multivariate statistics. J Med Chem. 2000;43:3867–77 (American Chemical Society).
36 Pathania S, Singh PK. Analyzing FDA-approved drugs for compliance of pharmacokinetic principles: should there be a critical screening parameter in drug designing protocols? Expert Opin Drug Metab Toxicol. 2021;17(4):351–4 (Taylor & Francis).
37. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem. 2002;45:2615–23 (American Chemical Society).

38 Goodsell DS, Zardecki C, Di Costanzo L, Duarte JM, Hudson BP, Persikova I, et al. RCSB Protein Data Bank: enabling biomedical research and drug discovery. Protein Sci. 2020;29(1):52–65 (John Wiley & Sons, Ltd).

39. Studio D. Dassault systemes BIOVIA, Discovery Studio modelling environment, Release 4.5. Accelrys Softw Inc. 2015;98–104.

40. Lawal B, Lee CY, Mokgautsi N, Sumitra MR, Khedkar H, Wu ATH, et al. mTOR/EGFR/iNOS/MAP2K1/FGFR/TGFB1 are druggable candidates for N-(2,4-difluorophenyl)-2′,4′-difluoro-4-hydroxybiphenyl-3-carboxamide (NSC765598), with consequent anticancer implications. Front Oncol. 2021;11:656738 (Frontiers Media S.A.).

41. Wu ATH, Lawal B, Wei L, Wen YT, Tzeng DTW, Lo WC. Multiomics identification of potential targets for Alzheimer disease and antrocin as a therapeutic candidate. Pharmaceutics. 2021;13:1555 (Multidisciplinary Digital Publishing Institute).

42. Lawal B, Liu YL, Mokgautsi N, Khedkar H, Sumitra MR, Wu ATH, et al. Pharmacoinformatics and preclinical studies of nsc765690 and nsc765599, potential stat3/cdk2/4/6 inhibitors with antitumor activities against nci60 human tumor cell lines. Biomedicines. 2021;9:1–22 (Multidisciplinary Digital Publishing Institute).

43. Lawal B, Kuo YC, Tang SL, Liu FC, Wu ATH, Lin HY, et al. Transcriptomic-based identification of the immuno-oncogenic signature of cholangiocarcinoma for hlc-018 multi-target therapy exploration. Cells. 2021;10:2873 (MDPI).

44. Lee JC, Wu ATH, Chen JH, Huang WY, Lawal B, Mokgautsi N, et al. Hnc0014, a multi-targeted small-molecule, inhibits head and neck squamous cell carcinoma by suppressing c-met/stat3/cd44/pd-l1 oncoimmune signature and eliciting antitumor immune responses. Cancers (Basel). 2020;12:1–18 (Multidisciplinary Digital Publishing Institute).

45 Sivajohan B, Elgendi M, Menon C, Allaire C, Yong P, Bedaiwy MA. Clinical use of artificial intelligence in endometriosis: a scoping review. NJP Digit Med. 2022;5(1):109.

46 Kimber-Trojnar Ż, Pilszyk A, Niebrzydowska M, Pilszyk Z, Ruszała M, Leszczyńska-Gorzelak B. The potential of non-invasive biomarkers for early diagnosis of asymptomatic patients with endometriosis. J Clin Med. 2021;10(13):2762 (Multidisciplinary Digital Publishing Institute).

47 Kvaskoff M, Mahamat-Sale Y, Farland LV, Shigesi N, Terry KL, Harris HR, et al. Endometriosis and cancer: a systematic review and meta-analysis. Hum Reprod Update. 2021;27(2):393–420 (Oxford Academic).

48 Vazgiourakis VM, Zervou MI, Papageorgiou L, Chaniotis D, Spandidos DA, Vlachakis D, et al. Association of endometriosis with cardiovascular disease: genetic aspects (review). Int J Mol Med. 2023;51(3):1–16 (Spandidos Publications).

49. Mori T, Ito F, Koshiba A, Kataoka H, Takaoka O, Okimura H, et al. Local estrogen formation and its regulation in endometriosis. Reprod Med Biol. 2019;18(4):305–11 (John Wiley & Sons, Ltd).

50 Tian Z, Chang XH, Zhao Y, Zhu HL. Current biomarkers for the detection of endometriosis. Chin Med J (Engl). 2020;133(19):2346–52 (Wolters Kluwer Health).

51 Li L, Sun B, Sun Y. Identification of functional TF-miRNA-hub gene regulatory network associated with ovarian endometriosis. Front Genet. 2022;13:998417.

52. Rekker K, Saare M, Roost AM, Kaart T, Sõritsa D, Karro H, et al. Circulating miR-200-family micro-RNAs have altered plasma levels in patients with endometriosis and vary with blood collection time. Fertil Steril. 2015;104:938-946.e2.

53. Kim BG, Yoo JY, Kim TH, Shin JH, Langenheim JF, Ferguson SD, et al. Aberrant activation of signal transducer and activator of transcription-3 (STAT3) signaling in endometriosis. Hum Reprod. 2015;30:1069–78.

54. Bianco B, Lerner TG, Trevisan CM, Cavalcanti V, Christofolini DM, Barbosa CP. The nuclear factor-kB functional promoter polymorphism is associated with endometriosis and infertility. Hum Immunol. 2012;73:1190–3.

55. Song Y, Xiao L, Fu J, Huang W, Wang Q, Zhang X, et al. Increased expression of the pluripotency markers sex-determining region Y-box 2 and Nanog homeobox in ovarian endometriosis. Reprod Biol Endocrinol. 2014;12:1–7.

56. Bernardi LA, Dyson MT, Tokunaga H, Sison C, Oral M, Robins JC, et al. The essential role of GATA6 in the activation of estrogen synthesis in endometriosis. Reprod Sci. 2019;26:60–9 (Society for Reproductive Investigation).

57 Izawa M, Taniguchi F, Harada T. GATA6 expression promoted by an active enhancer may become a molecular marker in endometriosis lesions. Am J Reprod Immunol. 2019;81:e13078 (John Wiley & Sons, Ltd).

58. Milewski Ł, Ścieżyńska A, Ponińska J, Soszyńska M, Barcz E, Roszkowski PI, et al. Endometriosis is associated with functional polymorphism in the promoter of heme oxygenase 1 (Hmox1) gene. Cells. 2021;10:1–11 (Multidisciplinary Digital Publishing Institute (MDPI)).

59. Chen P, Yao M, Fang T, Ye C, Du Y, Jin Y, et al. Identification of NFASC and CHL1 as two novel hub genes in endometriosis using integrated bioinformatic analysis and experimental verification. Pharmgenomics Pers Med. 2022;15:377–92.

60 Bae SJ, Jo Y, Cho MK, Jin JS, Kim JY, Shim J, et al. Identification and analysis of novel endometriosis biomarkers via integrative bioinformatics. Front Endocrinol (Lausanne). 2022;13:942368.

61. Dogan E, Saygili U, Posaci C, Tuna B, Caliskan S, Altunyurt S, et al. Regression of endometrial explants in rats treated with the cyclooxygenase-2 inhibitor rofecoxib. Fertil Steril. 2004;82:1115–20.

62. Kilico I, Kokcu A, Kefeli M, Kandemir B. Regression of experimentally induced endometriosis with a new selective cyclooxygenase-2 enzyme inhibitor. Gynecol Obstet. 2014;77:35–9 (Invest Karger Publishers).

63. Wang DB, Chen Q, Zhang C, Ren F, Li T. DNA hypomethylation of the COX-2 gene promoter is associated with upregulation of its mRNA expression in eutopic endometrium of endometriosis. Eur J Med Res. 2012;17:12.

64 Dolmans MM, Donnez J. Emerging drug targets for endometriosis. Biomolecules. 2022;12(11):1654.

65. Yamagata Y, Takaki E, Shinagawa M, Okada M, Jozaki K, Lee L, et al. Retinoic acid has the potential to suppress endometriosis development. J Ovarian Res. 2015;8(1):1–7 (BioMed Central).

66. Yilmaz BD, Bulun SE. Endometriosis and nuclear receptors. Hum Reprod Update. 2019;25:473–85.

67. Pavone ME, Dyson M, Reirstad S, Pearson E, Ishikawa H, Cheng YH, et al. Endometriosis expresses a molecular pattern consistent with decreased retinoid uptake, metabolism and action. Hum Reprod. 2011;26(8):2157–64 (Oxford University Press).