Computational and Structural Biotechnology Journal 21 (2023) 1510–1522

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Identification of *MKNK1* and *TOP3A* as ovarian endometriosis risk-associated genes using integrative genomic analyses and functional experiments

Yizhou Huang [a,1], Jie Luo [a,1], Yue Zhang [a], Tao Zhang [a], Xiangwei Fei [b], Liqing Chen [a], Yingfan Zhu [a], Songyue Li [a], Caiyun Zhou [c], Kaihong Xu [a], Yunlong Ma [d,*], Jun Lin [a,**], Jianhong Zhou [a,**]

[a] Department of Gynecology, Women's Hospital, Zhejiang University School of Medicine, Hangzhou 310006, Zhejiang Province, PR China
[b] Key Laboratory of Women's Reproductive Health of Zhejiang Province, Women's Hospital, Zhejiang University School of Medicine, Hangzhou 310006, Zhejiang Province, PR China
[c] Department of Pathology, Women's Hospital, Zhejiang University School of Medicine, Hangzhou 310006, Zhejiang Province, PR China
[d] Institute of Biomedical Big Data, School of Ophthalmology & Optometry and Eye Hospital, School of Biomedical Engineering, Wenzhou Medical University 325027 Wenzhou, Zhejiang Province, PR China

ARTICLE INFO

Article history:
Received 3 August 2022
Received in revised form 11 January 2023
Accepted 1 February 2023
Available online 5 February 2023

Keywords:
Endometriosis
Risk genes
Genome-wide association study
Expression quantitative trait loci
Integrative genomics analysis

ABSTRACT

The risk of endometriosis (EM), which is a common complex gynaecological disease, is related to genetic predisposition. However, it is unclear how genetic variants confer the risk of EM. Here, via *Sherlock* integrative analysis, we combined large-scale genome-wide association studies (GWAS) summary statistics on EM (N = 245,494) with a blood-based eQTL dataset (N = 1490) to identify EM risk-related genes. For validation, we leveraged two independent eQTL datasets (N = 769) for integration with the GWAS data. Thus, we prioritised 14 genes, including *GIMAP4*, *TOP3A*, and *NMNAT3*, which showed significant association with susceptibility to EM. We also utilised two independent methods, Multi-marker Analysis of GenoMic Annotation and S-PrediXcan, to further validate the EM risk-associated genes. Moreover, protein–protein interaction network analysis showed the 14 genes were functionally connected. Functional enrichment analyses further demonstrated that these genes were significantly enriched in metabolic and immune-related pathways. Differential gene expression analysis showed that in peripheral blood samples from patients with ovarian EM, *TOP3A*, *MKNK1*, *SIPA1L2*, and *NUCB1* were significantly upregulated, while *HOXB2*, *GIMAP5*, and *MGMT* were significantly downregulated compared with their expression levels in samples from the controls. Immunohistochemistry further confirmed the increased expression levels of *MKNK1* and *TOP3A* in the ectopic and eutopic endometrium compared to normal endometrium, while HOBX2 was downregulated in the endometrium of women with ovarian EM. Finally, in *ex vivo* functional experiments, *MKNK1* knockdown inhibited ectopic endometrial stromal cells (EESCs) migration and invasion. *TOP3A* knockdown inhibited EESCs proliferation, migration, and invasion, while promoting their apoptosis. Convergent lines of evidence suggested that *MKNK1* and *TOP3A* are novel EM risk-related genes.

1. Introduction

Endometriosis (EM), a common gynaecological disease characterised by the presence of endometrium-like tissue outside the uterus, affects approximately 10% of women of reproductive age [1]. EM presentation is highly heterogeneous, varying from cysts in the ovaries (endometrioma) and superficial peritoneal lesions to nodules with a depth of penetration exceeding 5 mm (deep EM) [1]. EM

* Correspondence to: Institute of Biomedical Big Data, School of Ophthalmology & Optometry and Eye Hospital, School of Biomedical Engineering, Wenzhou Medical University, Wenzhou 325027, Zhejiang Province, PR China.
** Correspondence to: The Department of Gynecology, Women's Hospital, Zhejiang University School of Medicine, 1 Xueshi Road, Hangzhou 310006, Zhejiang Province, PR China.
E-mail addresses: glb-biotech@zju.edu.cn (Y. Ma), linjun@zju.edu.cn (J. Lin), zhoujh1117@zju.edu.cn (J. Zhou).
[1] Yizhou Huang and Jie Luo contributed equally to this work.

https://doi.org/10.1016/j.csbj.2023.02.001

causes symptoms, including dysmenorrhoea, dyspareunia, chronic pelvic pain, and infertility, impairing women's quality of life while being a substantial economic burden [2,3]. Hormonal medication and surgical removal of lesions are the main therapeutic approaches for EM management; however, their efficacy is unsatisfactory, and EM recurrence rates are also high. A major reason for this dilemma is that the aetiology of this disease remains unclear.

Multiple lines of evidence suggest that EM is influenced by both genetic and environmental factors. Twin studies have demonstrated that EM heritability is approximately 50% [4], indicating the pivotal role of inherited risk variants in EM. From this perspective, a genome-wide association study (GWAS) is an effective approach for simultaneously examining the association of several single nucleotide polymorphisms (SNPs) with EM to the end of identifying novel risk genetic variants. The first GWAS, conducted in 2010 in the Japanese population, identified a significant association between two SNPs—rs10965235 located in CDKN2B-AS1 on chromosome 9 and rs16826658 located in WNT4 on chromosome 1—with EM [5]. Subsequently, more than 10 large-scale, population-based GWAS involving different ethnic populations have been performed, and a dozen genome-wide significant loci have been reported [6–8]. Nevertheless, as most reported risk variants are located in non-coding genomic regions, how these non-coding variants affect EM pathogenesis remains largely unknown. The greatest challenge in following up on GWAS is to identify genes responsible for an association with EM and to understand the functional consequences of these loci.

Expression quantitative trait loci (eQTL), which offers the possibility to elucidate a fraction of the genetic variance in gene expression, have been extensively studied for post-GWAS analyses [9–12]. Owing to its ability to establish a link between non-coding variants and the expression of a given gene, eQTL analysis is one of the most remarkable methods for highlighting disease-associated variants [13,14]. In this study, we performed a Bayesian integrative analysis (Sherlock) by combining genetic associations from large-scale GWAS summary statistics on EM and three independent eQTL datasets to identify potential EM risk-related genes. We further subjected peripheral blood samples from patients with EM and healthy controls to transcriptome sequencing to study the expression levels of the identified ovarian EM risk-associated genes. Additionally, the expression levels of these gens in tissue samples were investigated via immunohistochemical (IHC) analysis using ectopic, eutopic, and normal endometria samples. Finally, the potential roles of the EM risk-related genes were explored via in vitro functional studies. Our integrative study provided novel insights that *MKNK1* and *TOP3A* may represent promising EM risk genes.

2. Materials and methods

2.1. Multiple omics datasets

In the current integrative genomics analysis, we collected multi-omics datasets from several widely-established public databases as follows: 1) Dataset #1 for GWAS summary statistics on EM. We downloaded this GWAS summary dataset [15] from Gene ATLAS, an atlas of genetic associations in UK Biobank. A total of 245,494 subjects including 4252 EM patients based on European ancestry were chosen. The Affymetrix UK BiLEVE Axiom array and the Affymetrix UK Biobank Axiom array were used to genotype all samples.[16,17] After strict quality control, there were 13,853,045 SNPs eligible for downstream analyses. 2) Dataset #2 for GWAS data on Null phenotype: To evaluate the reliability of our findings, as referred to previous studies [18,19], we constructed a GWAS summary dataset on a randomly distributed phenotype (called as *Null* trait) as a negative control. [21,20,21] 3) Dataset #3 for discovery eQTL dataset. This eQTL dataset [22] as a discovery dataset was used to conduct the *Sherlock*-based integrative genomics analysis. There were 1490 subjects with a total of 675,350 SNPs and 12,808 genes. 4) Dataset #4 for validation eQTL dataset. This eQTL dataset [23] as an independent validation dataset was leveraged to perform the *Sherlock* inference analysis with the same parameters. A total of 400 subjects with 408,283 SNPs and 20,599 genes. 5) Dataset #5 for validation eQTL dataset. This blood-based eQTL dataset (n = 369 blood samples), which was downloaded from the Genotype-Tissue Expression (GTEx) portal (Version 7, https://gtexportal.org/) [24], was also adopted as an independent validation dataset for *Sherlock* analysis with the same parameters. For more detailed information for these datasets, please refer to Supplemental Methods.

2.2. Sherlock-based genomics analysis

We applied the *Sherlock* integrative genomics analysis [25] for incorporating GWAS summary statistics with eQTL data to uncover susceptible SNPs and genes for EM. The *Sherlock* Bayesian algorithm was designed to identify disease-relevant SNPs that have *cis*- and *trans*-regulatory effects on gene expression. The SNPs associated with EM and gene expression simultaneously were termed as eSNPs. There existed 3 potential scenarios for the Bayesian inference: 1) A positive score would be assigned based on an eSNP demonstrating a significant association with EM; 2) A negative score would be assigned based on an eSNP demonstrating a non-significant association with EM; 3) No score would be assigned based on an SNP that was significantly associated with EM but not associated with gene expression. The logarithm of the Bayes Factor (LBF) is calculated by summarising the assigned scores of all relevant eSNPs for a given gene as a vital indicator to predict EM-risk genes. The significance of *Sherlock* analysis for each gene is calibrated by a simulation analysis, and simulated *P*-value < 0.05 was of significance.

2.3. Gene-level genetic association analysis

The Multi-marker Analysis of GenoMic Annotation (MAGMA) [26] was used as an independent technique to conduct a genetic association analysis of the GWAS summary dataset on gene-level. In this updated model, MAGMA carry a T statistic:

$$T = \sum_i^N Z_i^2 = \boldsymbol{Z}^T\boldsymbol{Z}$$

where N is the SNP numbers within a specific gene. $Z_i = \varphi(p_i)$, φ represents the cumulative normal distribution function, and p_i is the marginal P value for a SNP *i*. SNPs were assigned to a given gene depended on the location of the SNP whether located into the gene or within a genomic region spanning 20 kb window of the gene. Gene body are defined as the region from transcription start site to transcription stop site. Furthermore, the MAGMA model assumes $\boldsymbol{Z} \sim \mathrm{MVN}(\boldsymbol{0}, \boldsymbol{S})$, where $\boldsymbol{S}$ is the linkage disequilibrium (LD) matrix among SNPs. The LD matrix can be diagonalized and thereby written as $\boldsymbol{S} = \boldsymbol{QAQ}^T$, where $\boldsymbol{Q}$ is an orthogonal matrix and $\boldsymbol{A} = \mathrm{diag}(\lambda_1, \lambda_2, ..\lambda_N)$ with λ_j being the *j*th eigenvalue of $\boldsymbol{S}$. The 1000 Genomes Project Phase 3 European Panel [17] is used as a reference for calculating LD among SNPs. $\boldsymbol{D} \sim \mathrm{MVN}(\boldsymbol{0}, \boldsymbol{I}_K)$ is a random variable, where $\boldsymbol{D} = \boldsymbol{A}^{-0.5}\boldsymbol{Q}^T\boldsymbol{Z}$. Thus, the sum of squared SNP Z-statistics is calculated:

$$T = \boldsymbol{Z}^T\boldsymbol{Z} = (\boldsymbol{QA^{0.5}D})^T\boldsymbol{QA^{0.5}D} = \boldsymbol{D}^T\boldsymbol{AD} = \sum_i^N \lambda_i D_i^2$$

where $D_i \sim \mathrm{N}(0, 1)$ and $D_i^2 \sim \chi_1^2$. T follows a mixture distribution of independent χ_1^2 random variables.

2.4. Integrating GWAS-based genetic association signals with eQTL data

To further validate the risk genes whose expression levels are linked with EM, we conducted an integrative genomics analysis by using the S-PrediXcan [27] as an independent approach to integrate meta-GWAS summary statistics with GTEx blood-based eQTL data (i.e., Dataset #5). S-PrediXcan mainly leverages two linear regression (MASHR) models to analyze the association between predicted gene expression and EM:

$$\boldsymbol{Y} = \alpha_1 + \boldsymbol{X}_l\beta_l + \varepsilon_1$$

$$\boldsymbol{Y} = \alpha_2 + \boldsymbol{G}_g\gamma_g + \varepsilon_2$$

where α_1 and α_2 are intercepts, ε_1 and ε_2 are independent error terms, $\boldsymbol{Y}$ is the n dimensional vector for n individuals, $\boldsymbol{X}_l$ is the allelic dosage for SNP l in n individuals, β_l is the effect size of SNP l, $G_g = \sum_{i\in gene(g)} \omega_{ig} X_i$ is the predicted expression calculated by ω_{lg} and $\boldsymbol{X}_l$, in which ω_{lg} is derived from the GTEx Project, and γ_g is the effect size of G_g. The Z-score (Wald-statistic) of the association between predicted gene expression and EM can be transformed as:

$$Z_g = \frac{\hat{\gamma}_g}{se(\hat{\gamma}_g)} \approx \sum_{i\in gene(g)} \omega_{ig} \frac{\hat{\sigma}_i}{\hat{\sigma}_g} \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$$

where $\hat{\sigma}_g$ is the standard deviation of G_g and can be calculated from the 1000 Genomes Project European Phase 3 Panel [17], $\hat{\beta}_l$ is the effect size from GWAS on EM and $\hat{\sigma}_l$ is the standard deviation of $\hat{\beta}_l$. Significant associations were adjusted using Bonferroni correction for multiple tests, and the significant P-value threshold was 4.46×10^{-6} (0.05/11,217).

2.5. Pathway-based enrichment analysis

To uncover the biological pathways and molecular functions of the identified EM-risk genes, we utilize the web-access tool of KOBAS [28] to performed a functional enrichment analysis (http://kobas.cbi.pku.edu.cn/kobas3). We submitted three gene sets identified by the *Sherlock* analysis (Gene sets #1–3) into the KOBAS website to calculate significantly enriched gene sets. There were two types of functional terms derived from multiple databases used in the enrichment analysis: 1) Biological pathways: Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, BioCyc, and PANTHER pathways; 2) Gene ontology (GO) terms: biological process, cellular component, and molecular functions. The hypergeometric test was used to calculate the significance of each functional term. The False discovery Rate (FDR) method was used for conducting multiple comparisons, and FDR < 0.05 was considered to be of significance.

2.6. Participants and clinical samples

The study participants were recruited in Women's Hospital, Zhejiang University School of Medicine from June 2020 to December 2020. Women who aged 18–45 years old and underwent hysterolaparoscopy for suspected ovarian EM were included in the ovarian EM group (N = 30). The peripheral blood samples were collected before surgery, and the eutopic and ectopic endometria were obtained simultaneously during surgery. For controls, the peripheral blood samples were collected from healthy women who had regular menstrual cycles (N = 30), and the normal endometria were obtained from women who underwent hysteroscopy for tubal infertility or uterine mediastinum (N = 30). All diagnosis were confirmed by surgery and final pathological examination. Subjects who had received hormonal treatment in the past 3 months, or with diseases of endocrine system, malignant tumour, major organ diseases or with pathological diagnosis of adenomyosis, polyps, fibroid, or endometrial hyperplasia were excluded. Participants' demographic and clinical characteristics are shown in Supplementary Table S1. All EM patients were classified as III/IV stage according to the revised American Fertility Society (r-AFS) classification [29]. Variables including age, BMI were not statistically significant between ovarian EM group and controls. The study was approved by the Human Ethics Committee of Women's Hospital, Zhejiang University School of Medicine (No. 20190014) and all participants signed informed consents.

The endometrium specimens were fixed for 24 h in 10% neutral buffered formalin then underwent routine processing of washing, dehydration, transparency, wax dipping, and embedding at the Department of Pathology of Women's Hospital, Zhejiang University School of Medicine for histologic diagnosis and immunohistochemistry. 5 mL peripheral blood was extracted with EDTA anticoagulant tube. Peripheral blood mononuclear cells (PBMCs) were isolated using a standardised density gradient technique.

2.7. RNA preparation and sequencing

Total RNA was extracted from the isolated PBMCs using the QIAzol and miRNeasy Mini Kit (Qiagen, CA, USA). After amplification, the RNA integrity was tested using the Bioanalyzer 2100 system with the RNA Nano 6000 Assay Kit (Agilent Technologies, CA, USA). The mRNA was purified from total RNA using poly-T oligo-attached magnetic beads, and used for establishing cDNA libraries for RNA sequencing. To preferentially select cDNA fragments of 370–420 bp in length, library fragments were purified using the AMPure XP system (Beckman Coulter, Beverly, USA). The quality of cDNA library was assessed on the Agilent Bioanalyzer 2100 system. Paired-end sequencing (150 bp) was performed on the Illumina Novaseq platform by Novogene (Beijing, China).

2.8. Transcriptomic mapping and profiling

First, raw sequencing data were qualified using the FastQC software (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The human reference genome file was downloaded from the Ensembl ftp site (http://asia.ensembl.org/info/data/ftp/index.html, file name: Homo_sapiens.GRCh37.75.cdna.all.fa). Index of the reference genome was established using Hisat2 V2.0.5 [30]. We used the Hisat2 V2.0.5 tool to align paired-end clean reads to the reference genome. The featureCounts V1.5.0-p3 [31] was utilised to count the reads number mapped to each gene. The Fragments Per Kilobase of exon model per million mapped fragments (FPKM) of each gene was calculated.

2.9. Differential gene expression analysis

To validate whether abnormal expression of these identified genetics-risk genes is associated with EM, we performed differential gene expression (DGE) analysis using the edgeR [32] R package (3.22.5) in our sequenced transcriptomic data that contained 30 EM patients and 30 matched controls. The *P*-values were corrected using the Benjamini & Hochberg method. The Student's t-test was applied to evaluate the significance.

Based on the Pearson correlation algorithm, we performed co-expression pattern analyses for discovering the co-expression patterns among the identified genes between EM and controls. The *Corrplot* package in R platform was used to visualise the co-expression patterns. To prioritise the important risk genes for subsequent functional validation, we performed an evidence scoring analysis for identified genes by combining all piece of supportive evidence from the current analysis including *Sherlock*, MAGMA, S-PrediXcan, and DGE analysis. A significant result scores 1, and a non-significant result scores 0.

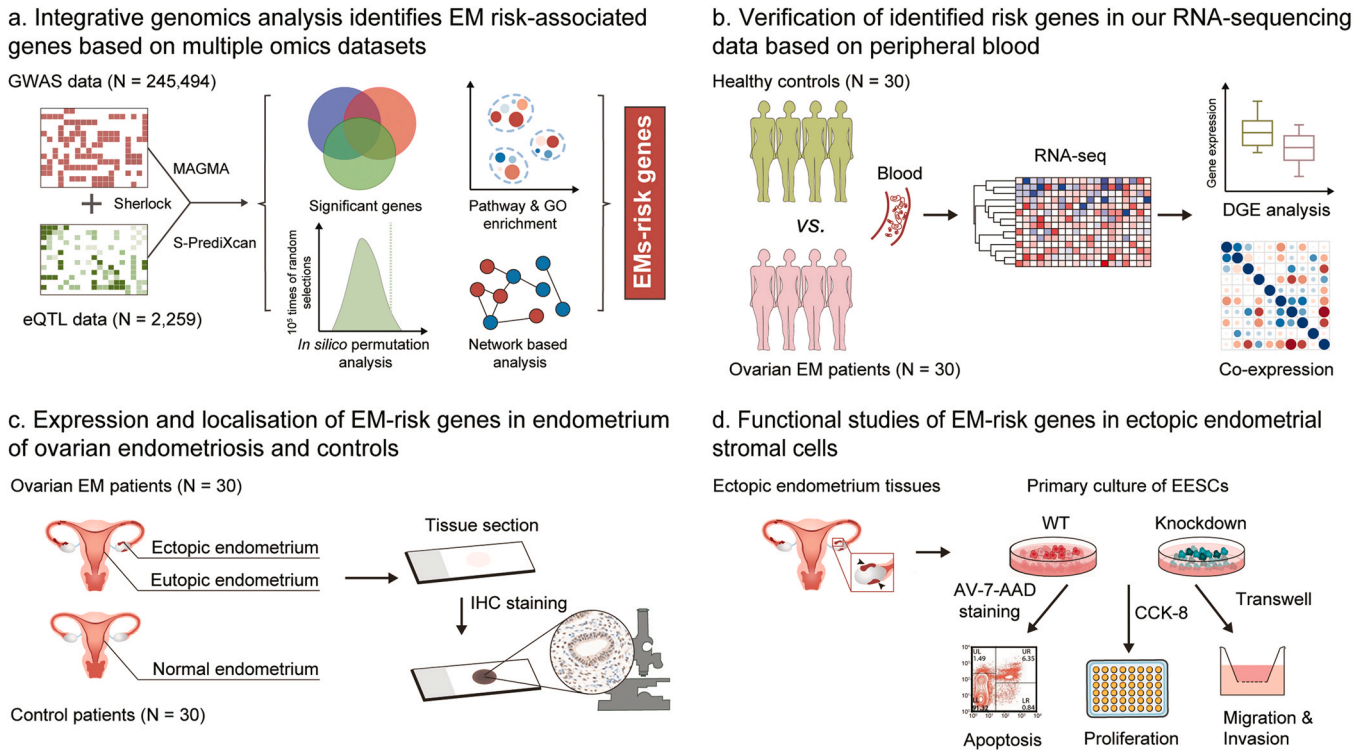

Fig. 1. Schematic of framework. EM, endometriosis. GWAS, genome-wide association study. eQTL, expression quantitative trait loci. MAGMA, Multi-marker Analysis of GenoMic Annotation. GO, Gene Ontology. DGE, differential gene expression. IHC, immunohistochemical. EESCs, ectopic endometrial stromal cells.

2.10. Immunohistochemistry staining and analysis

The IHC staining was performed as previously described[33]. 3 μm thick sections were cut from the tissue blocks, followed by routine deparaffinisation and rehydration procedures. IHC staining was performed with antibodies specific to MKNK1 (dilution 1:500, T611286; Abmart, Shanghai, China), TOP3A (dilution 1:1200, TD3265; Abmart, Shanghai, China), NUCB1 (dilution 1:2500, ab154262; Abcam, MA, USA) and HOXB2 (dilution 1:500, orb114161; Biorbyt, Cambridge, UK) for 1 h at room temperature. After washing, the sections were incubated with horseradish peroxidase (HRP)-conjugated secondary antibodies (Envision Detection kit, GeneTech, Shanghai, China) for 1 h and reacted with DAB (GeneTech) until appropriate for microscopic examination. A negative control was performed by the same method except for replacing the primary antibodies with PBS. Slides were evaluated independently by two blinded observers and re-examined by a senior pathology physician.

The IHC results were evaluated using the immunoreactive score (IRS) [34]. The percentage of positive cells was scored from 0 to 4 as: no, < 10%, 10–50%, 51–80%, and > 80%. The intensity of reaction was scored from 0 to 3 as: no colour, mild, moderate, and intense. The final score was calculated by multiplying the percentage and intensity scores, ranging from 0 to 12.

2.11. Small interfering RNA transfection in EESCs

Ectopic endometrial stromal cells (EESCs) were isolated and cultured as previously described [35]. Small interfering RNA (siRNA) were produced by Genepharma Corporation (Shanghai, China). The siRNA sequences were: MKNK1, 5'-GUGGGAUGAAACUGAACAATT-3′ (sense), 5'-UUGUUCAGUUUCAUCCCACTT-3′ (antisense); TOP3A 5'-GGCAGCAAGUGCAGAAAUATT-3′ (sense), 5'-UAUUUCUGCACUUGCUGCCTT-3′ (antisense). siRNAs (20 nM) were transfected into the EESCs at 70% confluency using Lipofectamine RNAiMAX (Invitrogen, Carlsbad, CA, USA). After transfection, total RNA (for 48 h) and protein (for 72 h) were extracted, and real-time quantitative PCR (RT-qPCR) and western blot analysis were conducted to assess transfection efficiency, respectively. For further details, see Supplementary Methods.

2.12. Biological behaviours assessment of EESCs

The biological behaviours of EESCs including proliferation, apoptosis, and migration and invasion were assessed via cell counting kit-8 assay, flow cytometry, and transwell assays, respectively. For detailed methods, see Supplementary Methods.

2.13. Statistical analysis

For clinical data, IHC scoring and cell experiments, statistical analyses were conducted using the SPSS 24.0 (IBM, USA) and GraphPad Prism 8 (GraphPad Software, USA). The continuous data were presented as mean ± SEM (or SD). Shapiro-Wilk test was used to examine the normality of data. For data variables with normal distribution, Student's t-test or one-way ANOVA followed by Bonferroni's post hoc tests were used for comparison between two groups or across multiple groups, respectively. For the data variables that were non-normally distributed, Mann–Whitney *U* or Kruskal–Wallis ANOVA followed by multiple comparison tests were carried out. The categorical data were shown as n (%) and compared by Chi-squared test. A value of $P < 0.05$ was indicated as statistically significant.

3. Results

3.1. Overview of the framework

In the current study, we leveraged a comprehensive integrative framework to prioritise novel EM susceptibility genes, based on multiple bioinformatics methods and functional experiments

(Fig. 1). To highlight the EM-associated risk genes, three steps were involved: 1) Using Sherlock integrative analysis, MAGMA, and S-PrediXcan methods, we combined GWAS summary statistics on EM (N = 245,494) with three independent eQTL datasets (N = 2259) to identify novel genes associated with EM risk and then used multiple bioinformatics approaches, including permutation, pathway-based enrichment, and PPI network analyses to examine the biological functions of the identified genes *in silico*. 2) To further validate the EM-risk genes, we collected peripheral blood samples from 30 patients with ovarian EM and 30 matched healthy controls for transcriptome sequencing and performed DGE and co-expression analyses. 3) To further explore the functional roles of the top-ranked newly identified EM-risk genes, we conducted follow-up validation studies by performing IHC and cellular functional experiments. Fig. 1 shows a detailed schematic of the proposed framework.

3.2. Prioritisation of EM risk-associated genes using Sherlock integrative analysis

The workflow of the integrative genomics analysis is shown in Supplementary Fig. S1. At the discovery stage, we integrated the GWAS summary statistics on EM (Dataset #1, N = 245,494) with eQTL data (Dataset #3, N = 1490) using *Sherlock* integrative analysis to determine whether abnormal gene expression confers susceptibility to EM. We found that 715 genes were significantly associated with EM risk (*Sherlock*-based permuted $P \leq 0.05$; Gene set #1, Supplementary Table S2). We re-performed the Sherlock analysis with the same parameter settings to validate the identified genes using two independent eQTL datasets (Datasets #4 and #5). In this regard, we found 683 significant EM risk-associated genes from Dataset #4 (*Sherlock*-based permuted $P < 0.05$; Gene set #2, Supplementary Table S3) and 330 significant EM risk-associated genes from Dataset #5 (*Sherlock*-based permuted $P < 0.05$, Gene set #3, Supplementary Table S4). Comparing the identified genes from Gene set #1 in the discovery stage with those from Gene sets #2 and #3, 14 overlapping genes across the three gene sets, including *GIMAP4* (permuted $P = 1.08 \times 10^{-3}$), *TOP3A* (permuted $P = 2.19 \times 10^{-3}$), and *NMNAT3* (permuted $P = 5.75 \times 10^{-3}$), were identified as EM risk-associated genes (Fig. 2A and Table 1).

Furthermore, we performed functional enrichment analyses for the three gene sets identified above based on the KEGG pathway and gene ontology (GO) terms using the KOBAS web-access tool. Thus, we observed that 19 common biological pathways and 35 common GO terms were significantly enriched by gene sets #1, #2, and #3 (FDR < 0.05, Fig. 2B–E and Supplementary Table S5). Based on Gene set #1, the top-ranked significantly enriched pathways included metabolism of proteins ($P = 1.27 \times 10^{-21}$) and immune system ($P = 4.66 \times 10^{-21}$), while the top-ranked significant GO term was membrane-bound organelles ($P = 1.76 \times 10^{-26}$).

3.3. Validation using in silico gene-level association analysis

To further validate the identified EM risk-associated genes, an independent method of gene-level genetic association analysis was adopted using the MAGMA tool. Thus, 1228 significant or suggestive genes associated with EM were found (MAGMA-based $P < 0.05$, Gene set #4, Supplementary Table S6). Consistently, 10 of the 14 *Sherlock*-identified risk genes were validated via the MAGMA analysis (Fig. 3A and Supplementary Table S7). Furthermore, none of the three *Sherlock*-identified EM risk-associated gene sets overlapped with the MAGMA-identified Null trait-related genes (Fig. 3B, Gene set #5, as a negative control).

To ensure the reliability of our findings, we performed *in silico* permutation analyses 100,000 times. Thus, we observed that the significant genes involved in Gene set #1 remarkably overlapped with Gene set #2 (Permuted $P = 2.0 \times 10^{-5}$; see Fig. 3C), Gene set #3 (Permuted $P = 0$; see Fig. 3D), and Gene set #4 (Permuted $P = 0$; see Fig. 3E), indicating that these identified EM-associated risk genes are biologically consistent. Moreover, to further ensure the reliability of our results, we compared the three identified significant gene sets based on *Sherlock* analysis with the gene sets based on the MAGMA analysis of GWAS summary statistics on EM and the Null phenotype. Thus, we observed that Gene sets #1, #2, and #3 showed markedly higher overlapping rates with Gene set #4 than with Gene set #5 at three different *P*-value thresholds (i.e., 0.05, 0.01, and 0.001; Supplementary Fig. S4), indicating that the identified genes associated with EM risk were attributable to genetic components instead of random events.

As referenced in previous studies [11,27,36], using the S-PrediXcan method as an independent technique to integrate GWAS summary statistics with GTEx blood-based eQTL data, we consistently found that nine of these 14 identified genes were associated with EM (Table 1). Also, we performed an integrative analysis of incorporating GWAS summary data on EM with GTEx eQTL data from uterus tissue that is more relevant tissue, and replicated six top-identified genes showing notable associations with EM ($P < 0.05$, Supplementary Table S8). To distinguish the causality of identified genes, we further conducted Mendelian randomization analyses of integrating GWAS summary data and two blood eQTL datasets from the eQTLGen and GTEx databases using SMR tool [13]. For GETx blood dataset, we found that nine of these 14 top-identified risk genes, including *TOP3A* and *MKNK1*, showed notable causality with EM ($P_{SMR} < 0.05$ and $P_{HEIDI} > 0.05$), and three genes exhibited suggestive causality with EM ($0.05 < P_{SMR} < 0.1$ and $P_{HEIDI} > 0.05$). As for eQTLGen blood dataset, we also identified nine genes showing notable causality with EM, including *TOP3A* and *MKNK1* ($P_{SMR} < 0.05$ and $P_{HEIDI} > 0.05$), and one gene exhibiting suggestively causal evidence for EM ($0.05 < P_{SMR} < 0.1$ and $P_{HEIDI} > 0.05$). Consistently, there is a high correlation of the SMR results between the eQTLGen and GTEx blood eQTL datasets (rho = 0.7, P = 0.0057, Supplementary Table S9 and Fig. S5). Notably, it has been reported that *TPM2*, *HOXB2*, and *MGMT* are associated with EM [37–39]. Based on our comprehensive integrative genomics analysis, we identified 14 genes, including 11 novel risk genes implicated in EM susceptibility (Table 1).

3.4. Network-based enrichment analysis of 14 EM risk-associated genes

To investigate the underlying molecular links corresponding to the 14 EM risk-associated genes, we constructed PPI network analysis using the GeneMANIA tool. From Fig. 4, it is evident that strong biological interactions existed among these identified risk genes; this is consistent with previous evidence that biologically related genes may demonstrate convergent contribution to complex disease risk [40,41]. Co-expression links accounted for most (71.51%) of the molecular interactions among these identified risk genes. For example, *NUCB1* showed co-expression with *TPM2*, *HOXB2*, and *MGMT*. Additionally, protein domains were also shared among these genes. Notably, these protein domains accounted for 27.25% of the total network. Our PPI network analysis also demonstrated that the 14 identified EM risk-associated genes might have a synergistic contribution to the pathogenesis of EM.

3.5. Verification of identified EM-risk genes in our RNA-sequencing data based on peripheral blood

To further validate these results, we subjected peripheral blood samples from 30 ovarian EM cases and 30 healthy controls to DGE analysis via RNA sequencing. The expression levels of the 14 EM risk-associated genes in each individual were visualized as heatmaps as shown in Supplementary Fig. S6. Further, we examined the differences in the expression levels of the 14 EM risk-associated genes identified by Sherlock analysis between the groups. As

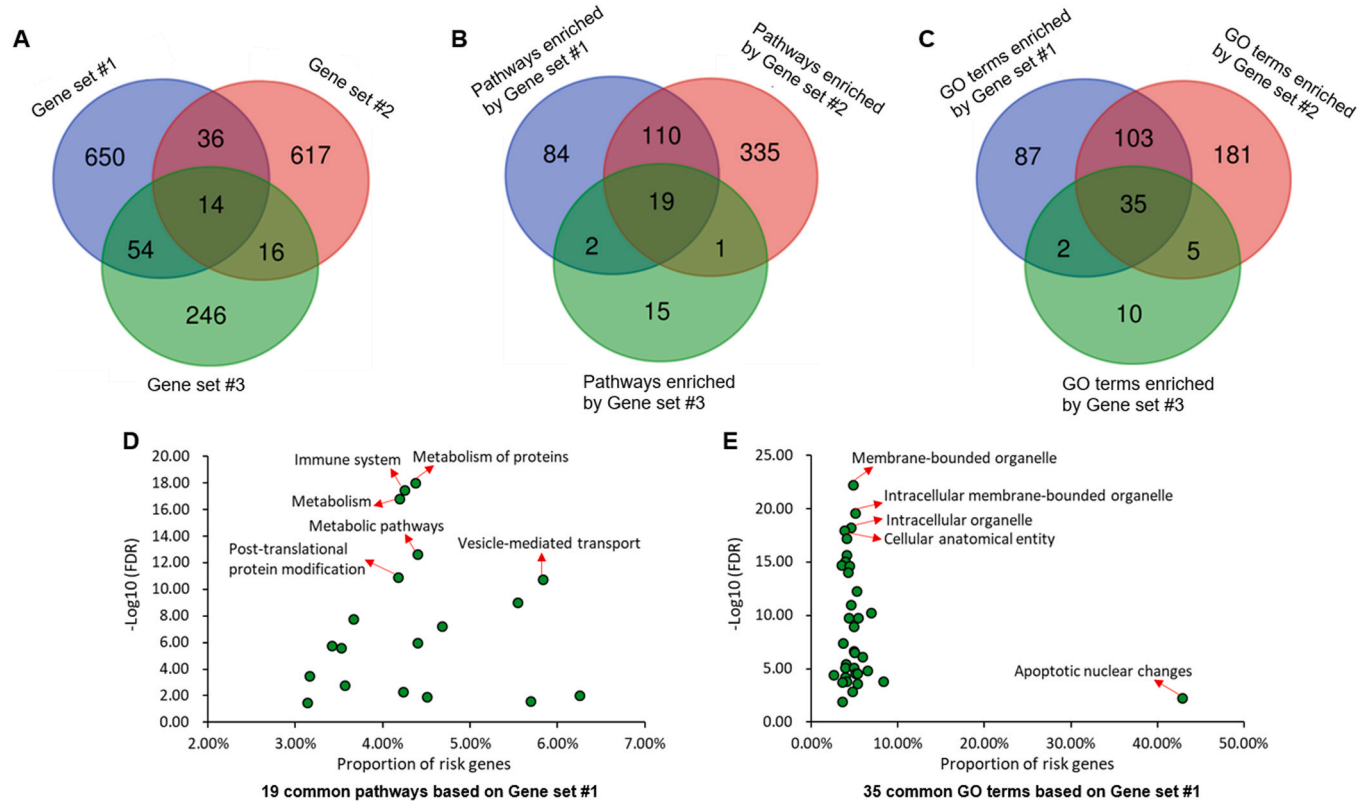

Fig. 2. Integrative genomics analyses identify risk genes and pathways for EM. (A) Venn diagram of three identified EM-risk gene sets: Gene set #1, Gene set #2, and Gene set #3 are based on Sherlock integrative genomics analysis by combining Zeller et al. eQTL data (Dataset #3), Dixon et al. eQTL data (Dataset #4), and GTEx blood eQTL data (Dataset #5) with GWAS summary statistics on EM, respectively. (B, C) Venn diagrams of the significantly enriched pathways (B) and GO terms (C) by three identified gene sets. (D, E) The scatter diagrams showing the 19 common significant pathways (D) and 35 common significant GO terms (E) based on Gene set #1. EM, endometriosis. GO, Gene Ontology.

shown in Fig. 5A–G and Table 2, among the 14 genes, *TOP3A*, *MKNK1*, *SIPA1L2*, and *NUCB1* were significantly upregulated, while *HOXB2*, *GIMAP5*, and *MGMT* were significantly downregulated in PBMC from patients with ovarian EM compared with the samples from the controls. The changes in the expression levels of the other seven genes (*GIMAP4*, *NMNAT3*, *TPM2*, *METTL27*, *VAMP4*, *ENDOG*, and *RBM18*) did not show any significant difference between the EM cases and the healthy controls (Supplementary Fig. S7). By performing co-expression pattern analysis, we found that the co-expression patterns of the 14 important genes were prominently altered among the patients with EM compared with the control cases (Fig. 5H). Overall, our RNA sequencing results provided further evidence that the identified genes showed aberrant expression levels in EM compared with the healthy controls. Based on the scoring of the multiple supportive evidence corresponding to the 14 genes (Table 2), we selected the top five genes, *TOP3A*, *MKNK1*, *SIPA1L2*, *NUCB1*, and *HOXB2*, for follow-up protein endometrial tissue expression evaluation.

3.6. *MKNK1* and *TOP3A* are highly expressed in the endometrium of women with ovarian EM

To evaluate the protein expression levels and localisation of the identified EM risk-associated genes in the endometrium of the patients with EM and the healthy controls, we performed IHC staining for *TOP3A*, *MKNK1*, *SIPA1L2*, *NUCB1*, and *HOXB2* in ectopic

Table 1
Comprehensive genomics analyses showing that 14 genes are implicated in EM risk.

Gene	LBF in the discovery stage	Permuted *P*-value (*Sherlock* analysis of Dataset #3)	Permuted *P*-value (*Sherlock* analysis of Dataset #4)	Permuted *P*-value (*Sherlock* analysis of Dataset #5)	MAGMA-based *P*-value (Dataset #1)	MAGMA-based *P*-value (Dataset #2, Negative control)	S-PrediXcan-based *P*-value (GTEx v7 whole blood)
GIMAP4	3.41	0.0011	0.0023	0.0019	0.0014	0.13	0.0049
TOP3A	2.86	0.0022	0.024	0.0028	0.011	0.11	0.0032
NMNAT3	2.12	0.0058	0.006	0.020	0.12	0.52	0.23
MKNK1	1.70	0.0098	0.032	0.0039	3.88E-05	0.65	0.027
TPM2	1.62	0.011	0.034	0.012	0.025	0.57	0.0015
SIPA1L2	1.57	0.012	0.041	0.019	0.32	0.080	0.0017
METTL27	1.44	0.014	0.011	0.0087	0.0046	NA	0.048
NUCB1	1.26	0.017	0.023	0.0098	0.0079	0.35	0.0081
VAMP4	1.22	0.018	0.04	0.0082	0.013	0.38	0.0099
ENDOG	1.01	0.024	0.041	0.026	0.026	0.22	0.056
HOXB2	0.85	0.029	0.04	0.0088	0.051	NA	0.018
GIMAP5	0.57	0.044	0.0063	0.048	0.0014	0.62	0.46
RBM18	0.55	0.045	0.0078	0.040	0.12	0.80	0.84
MGMT	0.54	0.046	0.023	0.036	0.0052	0.90	0.84

EM, endometriosis; LBF, logarithm of the Bayes Factor; MAGMA, Multi-marker Analysis of GenoMic Annotation.

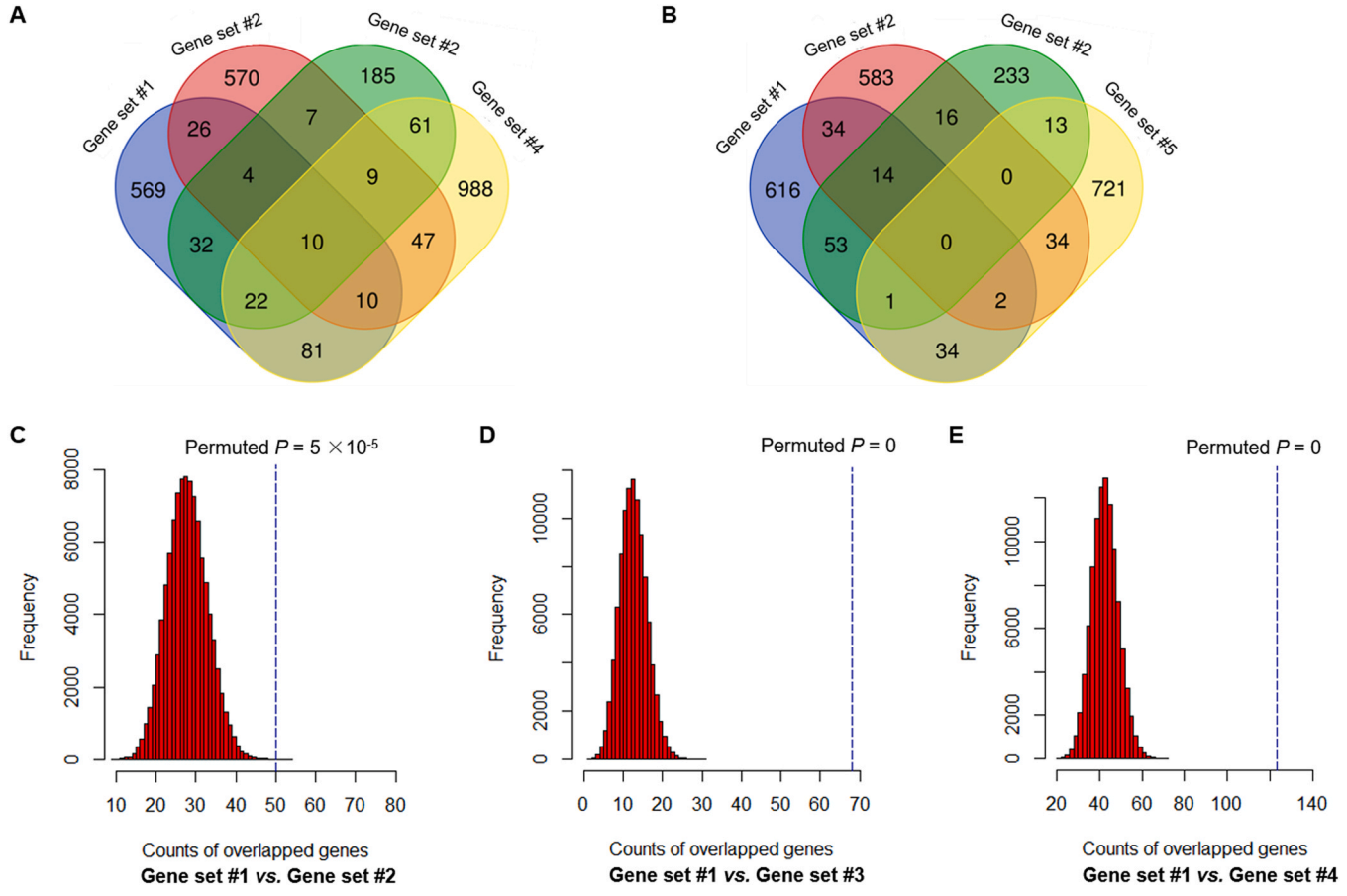

Fig. 3. Consistent evidence supporting the identified EM-risk genes from integrative genomics analyses. (A, B) Venn diagrams showing the overlapping genes of three Sherlock-identified EM-risk gene sets with MAGMA-identified EM-risk genes (Gene set #4; A), and with MAGMA-identified Null trait-related genes (Gene set #5, as a negative control; B). (C–E) Computer-based permutation analyses of 100,000 times for comparison of genes from Gene set #1 with that from Gene set #2 (C), Gene set #3 (D), and Gene set #4 (E). EM, endometriosis. MAGMA, Multi-marker Analysis of GenoMic Annotation.

and eutopic endometria samples from 30 patients with ovarian EM and normal endometria samples from the 30 control patients. Thus, we observed that *MKNK1* was primarily localised in the nucleus of endometrial glandular epithelial cells, whereas its expression in the endometrial stroma was comparatively weak (Fig. 6A–C). Additionally, *MKNK1* expression was significantly higher in eutopic endometrium than in normal endometrium and showed the highest expression level in ectopic endometrium ($P < 0.05$; Fig. 6D). Further, *TOP3A* was predominantly immunolocalised in the cytomembrane and cytoplasm of endometrial glandular epithelial cells (Fig. 6E–G). Additionally, its expression level (IRS) in the ectopic endometrium was significantly higher than that in the eutopic and normal endometrium ($P < 0.001$ for both), and its expression in the eutopic endometrium was increased compared with that corresponding to normal endometrium ($P < 0.01$), as presented in Fig. 6H. Endometrial glandular epithelial and stromal cells expressed *HOXB2*, and the protein of this gene was primarily localised in the cytoplasm and nucleus (Fig. 6I–K). Additionally, ectopic endometrium showed decreased *HOXB2* expression compared with eutopic and normal endometrium ($P < 0.01$), but the latter two showed no significant difference in this regard ($P > 0.05$), as can be seen in Fig. 6L. *NUCB1* was expressed in both the endometrial stroma and epithelium and mainly presented the glandular epithelial cells cytoplasmic staining in (Fig. 6M–O). Ectopic, eutopic, and normal endometria did not differ significantly with respect to *NUCB1* expression ($P > 0.05$, Fig. 6P). Unfortunately, due to the lack of a good antibody, *SIPA1L2* expression could not be detected (data not shown).

3.7. *MKNK1* and *TOP3A* regulate the biological behaviours of EESCs

As *MKNK1* and *TOP3A* were identified as novel, top-ranked EM-risk genes, we investigated the functional roles of these two genes in EM in vitro. EESCs were transfected with *MKNK1* siRNA and *TOP3A* siRNA to knock down their expression, and their biological behaviours were assessed. The results of RT-qPCR and western blotting showed significantly lower *MKNK1* and *TOP3A* expression in siRNA-transfected EESCs (Fig. 7A–C). Further, the results of the CCK-8 assay showed a significant decrease in the proliferation of EESCs in the si-*TOP3A* group, while no difference was observed in the si-*MKNK1* group compared with the si-Ctrl group (Fig. 7D). Based on flow cytometry, the apoptosis rate corresponding to the si-*TOP3A* group was higher than that corresponding to the si-Ctrl group (Fig. 7E). Furthermore, based on Transwell assays, compared with the si-Ctrl group, the si-MKNK1 and si-TOP3A groups showed significantly impaired EESC migration and invasion abilities (Fig. 7F). Our cell function studies also indicated that *MKNK1* downregulation inhibited the migration and invasion abilities of EESCs, but did not affect their proliferation and apoptosis rates. Additionally, *TOP3A* downregulation inhibited EESCs proliferation, migration, and invasion and promoted their apoptosis.

4. Discussion

EM is a common and complex disease with genetic predisposition. Hitherto, multiple GWAS have been performed to reveal the genetic determinants underlying EM in populations worldwide.

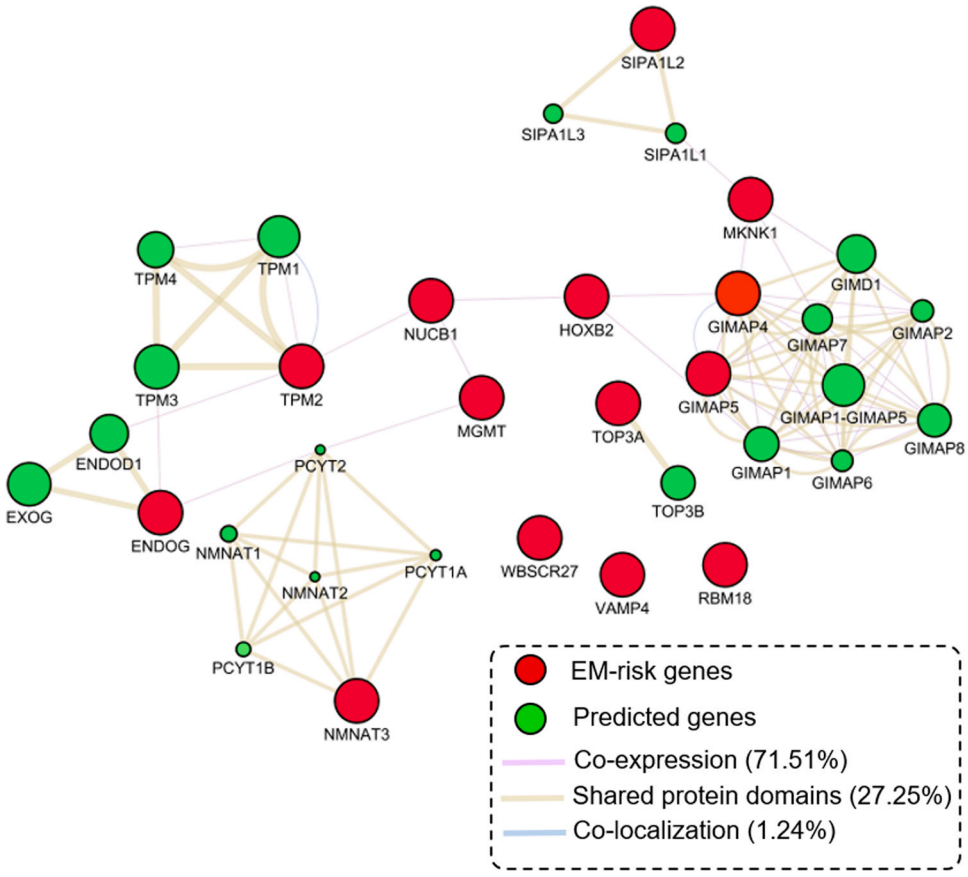

Fig. 4. PPI-network of 14 EM-risk genes using the GeneMANIA tool. The 14 EM-risk genes were marked with red colour, and the 20 predicted genes were marked with green colour. The underlying molecular interactions between each gene pair were attributed based on the co-expression links (account for 71.51%), shared protein domains (account for 27.25%), and co-localization (account for 1.24%). PPI, protein–protein interaction.

However, the genes associated with EM risk have not yet been identified. As most of the EM risk-associated risk variants are located in non-coding regions, the identified risk variants may confer the risk of EM by regulating gene expression. Therefore, identifying risk genes, from our perspective, is a crucial step in bridging GWAS findings and the aetiology of EM to the end of facilitating the development of novel therapeutics for its management. In this study, we conducted *Sherlock* analyses to integrate a large-scale EM GWAS dataset with independent eQTL datasets. Thus, we first identified 14 risk genes whose expression changes may contribute to the risk of EM and thereafter, performed comprehensive analyses to validate and prioritise the identified risk genes. Further DGE analysis showed that seven of these genes, including *TOP3A*, *MKNK1*, and *SIPA1L2* were dysregulated in the peripheral blood of ovarian EM cases compared with the control samples. IHC staining results also consistently showed that *MKNK*1 and *TOP3A* were upregulated, while *HOBX2* was downregulated in the endometrium of women with ovarian EM. Finally, we observed that the knockdown of *MKNK1* and *TOP3A* affected the migration and invasion behaviours of EESCs. Taken together, these convergent lines of evidence suggested that *MKNK1* and *TOP3A* are promising candidate genes for EM.

MKNK1 (also named *MNK1*), located on chromosome 1p33, plays essential roles in many human diseases, including tumourigenesis and metabolic diseases, and is also implicated in autoimmune and inflammatory diseases as well as viral replication processes. Additionally, *MKNK1* is one of the immediate downstream effectors of the activated MAPK and PI3K pathways driven by $BRAF^{V600E}$ and mutated *PTEN*. Elevated levels of MKNK protein kinases and their substrate, eIF4E (or p-eIF4E), have been detected in multiple types of solid tumours (e.g., breast, prostate, and melanoma) as well as haematological malignancies [42–44]. It has also been reported that MKNK1 facilitates tumour invasion and metastasis by promoting eIF4E phosphorylation [45,46]. In addition, numerous studies have highlighted that the MKNK/eIF4E axis contributes to promoting oncogenic mRNA translation [44,47], and in recent years, MKNK1/2 has been regarded as an important molecular target in invasive and metastatic cancer, and several MKNK1/2 inhibitors have reached phase I/II clinical trials [42]. However, despite the vital role of the MAPK and PI3K pathways in EM, no study to date has focused on the association between MKNK1 with EM. In this study, we identified *MKNK1* as a promising EM risk-related gene and verified that it was consistently upregulated in peripheral blood and endometrium samples from EM cases compared with controls. Consistent with its known cellular function, *MKNK1* protein expression was detected in both the nucleus and cytoplasm of endometrial cells. Our in vitro experiments also suggested that *MKNK1* possibly participates in the pathogenesis of EM by promoting the invasion and migration of EESCs, the mechanism of which might involve eIF4E phosphorylation or the regulation of other oncogenic cell signalling pathways.

TOP3A is located on chromosome 17p11.2 and encodes Top3α (topoisomerase 3α), a type IA DNA topoisomerase that shows dual localisation, in the nucleus and mitochondria [48]. Reportedly, the nuclear isoform of Top3α functions as a decatenase, facilitating the processing of homologous recombination intermediates to maintain genomic stability [49,50]. Additionally, the mitochondrial isoform of Top3α is an essential component of the mtDNA replication machinery required for the decatenation and segregation process [51]. However, the function of *TOP3A* in human diseases, including EM, has not yet been sufficiently investigated. Although the pathophysiology of EM remains elusive, dysregulated DNA damage response

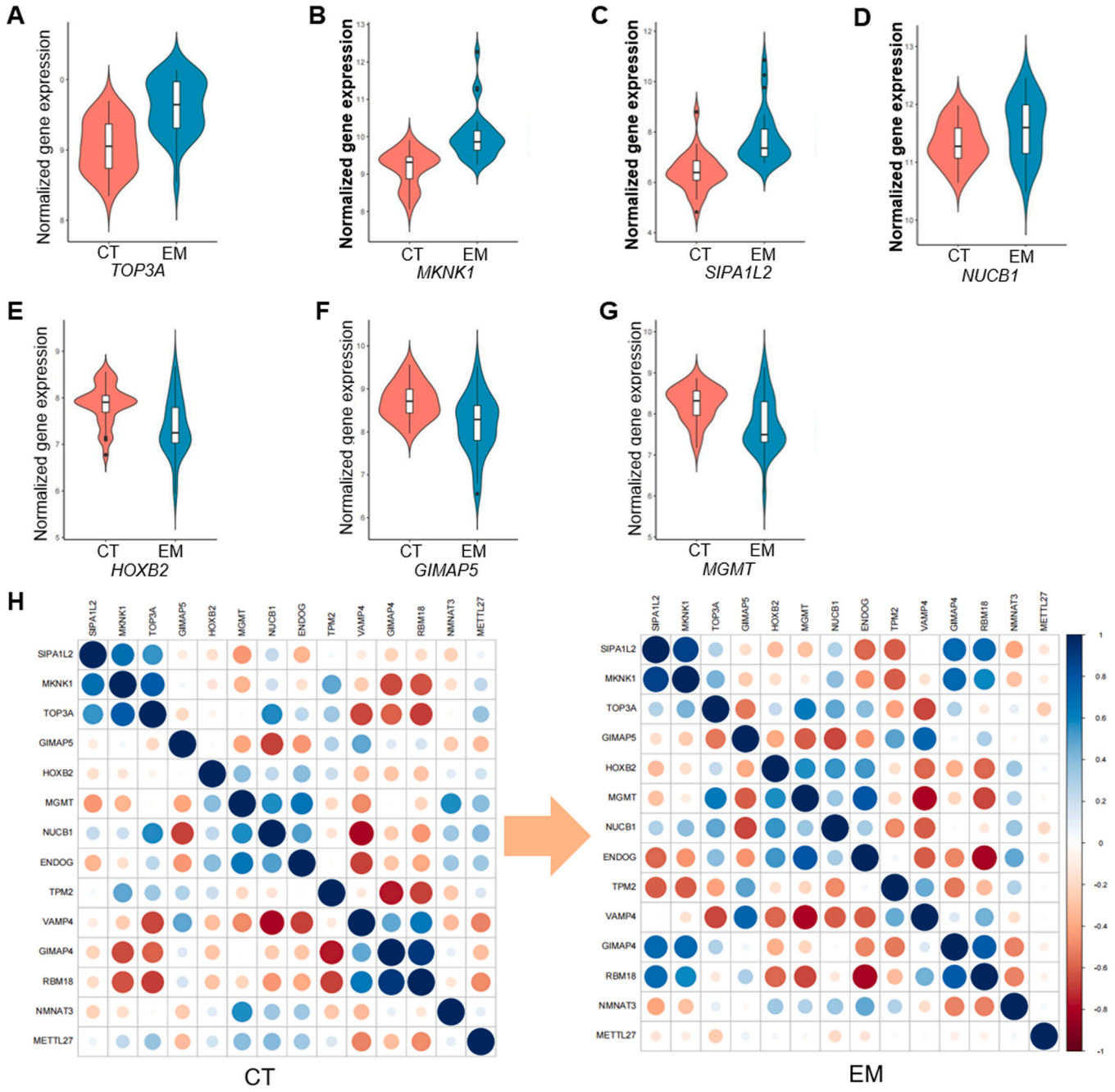

Fig. 5. Differential gene expression and co-expression patterns of EM-risk genes in PBMCs of women with ovarian EM and health controls based on RNA-Seq. (**A-G**) Violin plots showing significantly different expressed genes between EM and controls for *TOP3A* (A), *MKNK1* (B), *SIPA1L2* (C), *NUCB1* (D), *HOXB2* (E), *GIMAP5* (F), and *MGMT* (G). (**H**) Co-expression pattern analysis of 14 EM-risk genes between controls and EM. The colour legend showing the degree of correlation coefficients, red represents − 1 and blue represents + 1. PBMC, peripheral blood mononuclear cell. EM, endometriosis.

(DDR) has received much attention in this regard in recent years. For example, Bane et al. [52] demonstrated that eutopic endometrium from women with EM show higher DDR and DNA repair gene expression levels, as well as higher DNA damage levels compared with the controls, suggesting the existence of stimuli that induce DNA damage in eutopic endometrium. Thus, the involvement of *TOP3A* in homologous recombination repair may provide clues regarding its biological role in EM. High TOP3A expression levels in eutopic and ectopic endometrium samples probably help counteract the high DNA damage caused by external or internal factors; notwithstanding, this still warrants further investigation.

HOXB2, a member of the HOX family, is a transcription factor that is involved in embryonic development. The expression of HOXB2 is altered in a variety of solid tumours, the function of which could be distinct in different tumours. HOXB2 was identified as a tumor suppressor in breast cancer cells, whose expression could be downregulated by estrogen[53]. A previous study demonstrated that HOXB2, as a downstream target of miR-202–5, played a role in inhibiting the proliferation, invasion and migration of ovarian cancer cells[54]. However, in some other malignant tumours such as esophageal squamous cell carcinoma, nephroblastoma and bladder cancer, HOXB2 presented a tumour promotor via increasing cell proliferation, invasion and migration [55–57]. In EM, whether HOXB2 act as a suppressor thus the decreased expression of protein might promote diseases progression need to be further clarified.

Table 2
Seven significantly differentially expressed genes verified by subjecting PBMCs from women with ovarian EM and healthy controls to RNA-sequencing.

Gene names	Log2 Fold Change	*P*-value	Evidence Scoring
TOP3A	0.56	1.15E-08	6
MKNK1	0.94	1.91E-09	6
NUCB1	0.30	0.011	6
SIPA1L2	1.62	1.27E-10	5
HOXB2	-0.42	0.0021	5
GIMAP5	-0.46	0.0013	5
MGMT	-0.42	0.003	5

Note: Evidence scores were calculated by combining all pieces of supportive evidence from the analyses performed in this study, including *Sherlock*, MAGMA, S-PrediXcan, and DGE analyses. A score of 1 indicates a significant result, while a score of 0 indicates a non-significant result. EM, endometriosis; PBMC, peripheral blood mononuclear cell; MAGMA, Multi-marker Analysis of GenoMic Annotation; DGE, differential gene expression.

Previous GWASs have been conducted for identifying disease-associated variants for EM [58]. To give an overview of significant loci obtained through GWAS, we summarized these reported significant variants in the Supplementary Table S10. Among the 14 identified genes, *TPM2*, *HOXB2*, and *MGMT* have been shown to be associated with EM in a few previous studies [37–39]. Specifically, *TPM2* encodes beta-tropomyosin, which plays a role in muscle contraction and motility, and helps maintain cell shape and cell-matrix interactions. Irungu et al. [37] discovered and confirmed that *TPM2* is highly expressed in the ectopic endometrium and serum of patients with EM compared with samples from the controls, suggesting it has potential as a biomarker of EM. In this study, bioinformatics analyses based on public datasets suggested that the expression level of *TPM2* is associated with EM risk, but showed no expression difference in our verification samples. This phenomenon could be attribute to several reasons including the different ethnic background, relatively small verification sample size, and differential expression of genes in different tissues and different levels. Vestergaard et al. [38] observed that HOXB2 transcription was significantly reduced in EM lesions compared with endometrium samples from patients with EM and healthy controls, which was in agreement with our results. O6-methylguanine-DNA methyltransferase (MGMT), which is responsible for the direct repair of DNA, is primarily immunolocalised in the nuclei of epithelial cells in eutopic endometrial tissue and ovarian EM lesions [39]. Nevertheless, previous studies have only been focused on the investigation of the expression or localisation patterns of risk genes. Thus, further studies are needed to clarify the function of these genes in EM.

Several eQTL analysis studies have focused on the association between genetic variation and gene expression in EM. In this regard, Montgomery et al. have highlighted three eQTLs that may regulate the expression of target genes, including *LINC0039* and *CDC42* on chromosome 1, *VEZT* on chromosome 12, and *CDKN2A-AS1* on chromosome 9, by integrating gene expression data from whole blood (n = 862) and endometrial tissue (n = 136) with an Australian GWAS dataset (2594 cases and 4496 controls) [59,60]. Recently, Chou et al. conducted a GWAS involving 126 EM cases and 96 controls in a Taiwanese population, and thereafter, mapped the results obtained with the GTEx database. They identified that SNP rs13126673 on chromosome 4 is a cis-eQTL and is associated with both EM risk and *INTU* expression [61]. These previous studies were based on relatively small sample sizes of GWAS data, which may have reduced their power to identify more risk loci. In this current

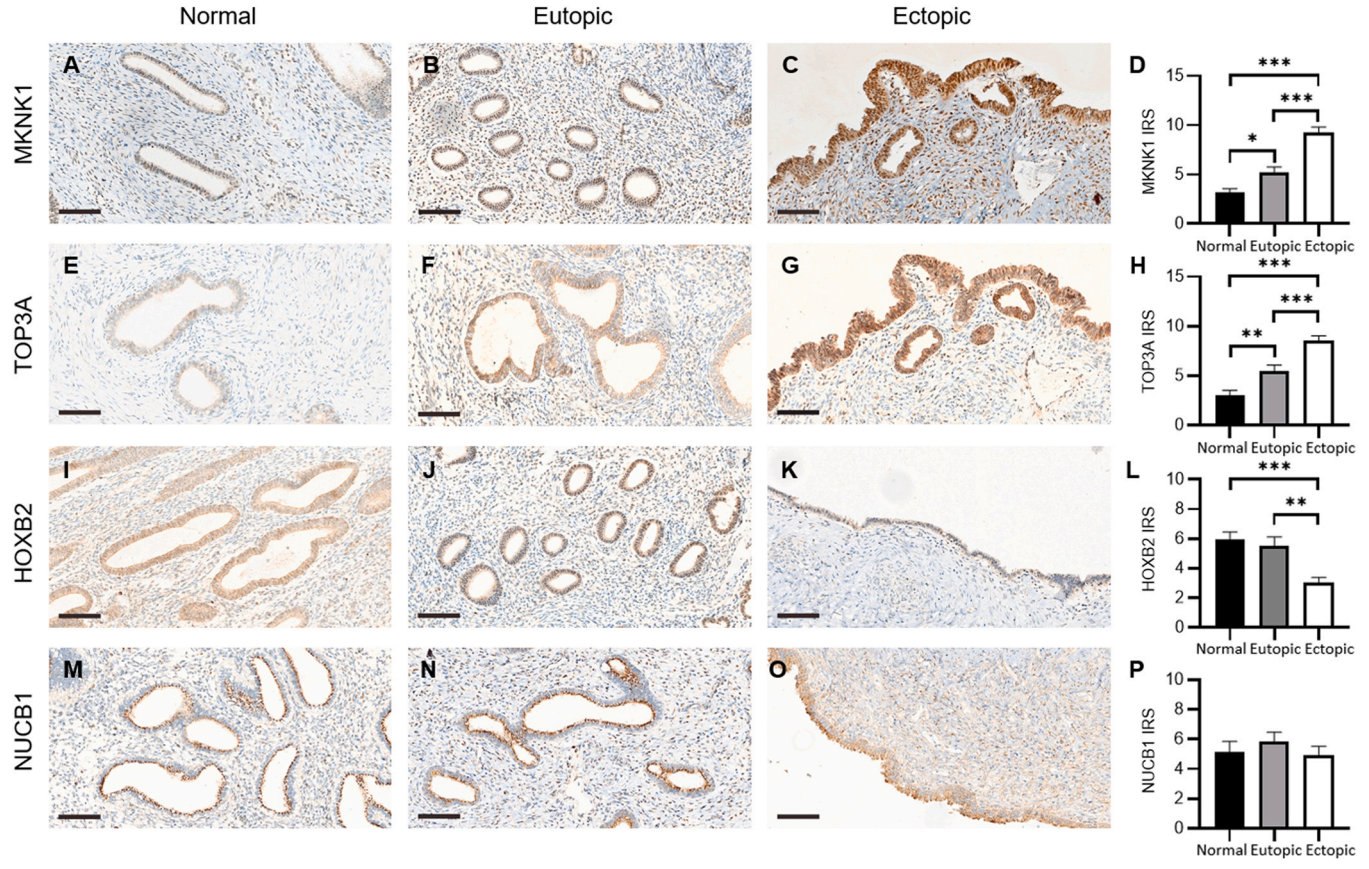

Fig. 6. Immunoreactivity of MKNK1, TOP3A, HOXB2 and NUCB1 in endometrium from women with and without ovarian endometriosis. The expression and localisation of MKNK1 (A–C), TOP3A (E–G), HOXB2 (I–K), NUCB1 (M–O) in normal, eutopic, and ectopic endometrium evaluated by IHC staining, respectively. The comparisons of IRS across three groups (D, H, L, P). Values are presented as mean ± SEM. *P*-values were determined by Kruskal-Wallis tests followed by multiple comparisons. $*P < 0.05$, $**P < 0.01$. and $***P < 0.001$. IHC, immunohistochemical. IRS, immunoreactive score.

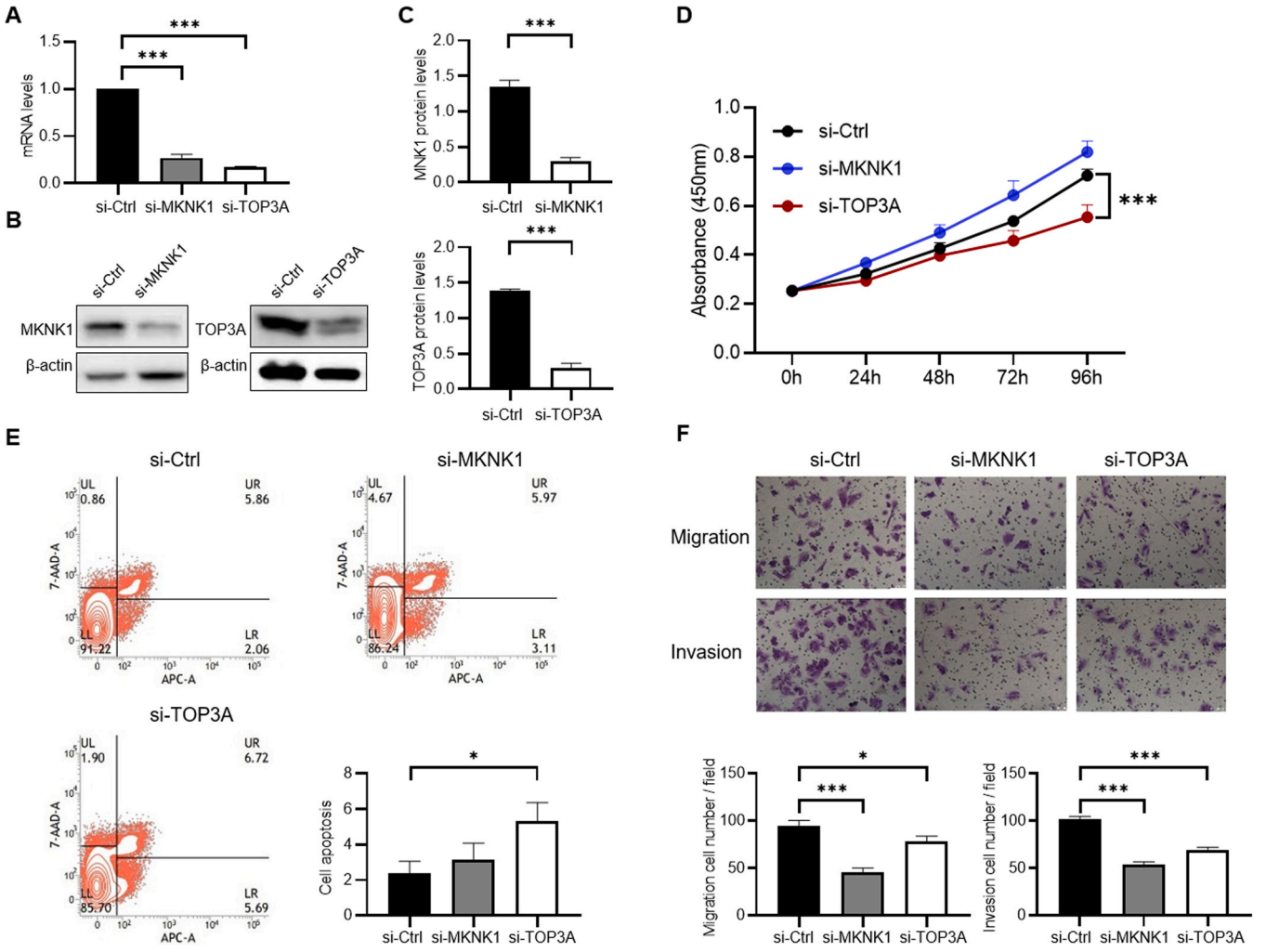

Fig. 7. The role of MKNK1 and TOP3A in proliferation, apoptosis, migration and invasion of EESCs. (A–C) mRNA (A) and protein (B–C) expression levels of MKNK1 and TOP3A in EESCs transfected with siRNA were determined by RT-qPCR (n = 3) and western blot analysis (n = 6), respectively. (D) Proliferation of EESCs transfected with si-MKNK1, si-TOP3A, and si-Ctrl was assessed with CCK-8 assay at 0, 24, 48, 72 and 96 h, n = 6. (E) Representative images and the graphical statistics of apoptosis rate assessed by flow cytometry of EESCs transfected with si-MKNK1, si-TOP3A, and si-Ctrl. n = 6. (F) Representative fields (100 × magnification) and the graphical statistics of Transwell migration and invasion assay of EESCs transfected with si-MKNK1, si-TOP3A, and si-Ctrl. n = 5. Values are presented as mean ± SEM. *P*-values were determined by unpaired two-tailed t test. $*P < 0.05$, and $***P < 0.001$. EESCs, ectopic endometrial stromal cells. Real-time quantitative PCR (RT-qPCR). Cell counting kit-8, CCK-8.

study, we leveraged GWAS data with a very large sample size (n = 245,494) from the UK Biobank database and three independent eQTL datasets for integrative genomic analysis; this enhanced the possibility of identifying more novel loci. Additionally, *Sherlock* integrative genomics analysis, based on the Bayesian inference method, is a vigorous tool for integrating genetic data from GWAS with existing eQTL data [25]. Compared with the usual GWAS approaches that disregard large amounts of common genetic variants with minor effects, Sherlock integrative analysis has an obvious advantage in that it involves the re-use of these disregarded common variants in GWAS. Further, based on *Sherlock* analysis, several novel risk genes have been implicated in the pathogenesis of various complex diseases, including schizophrenia, childhood-onset asthma, major depressive disorders, and COVID-19 [18,36,62,63]. Our study further provides supportive evidence that incorporating multiple layers of omics data contributes to strengthening the association signals of pinpointing risk loci for complex diseases.

This study had several limitations that should be considered. First, the GWAS data and eQTL data used in the integrative analysis were based on European ancestry, whereas our RNA-sequencing data were derived from a Han Chinese population. This might have led to biases due to the differences in genetic architectures across different ethnicities. Second, even though our current integrative genomic analyses have highlighted some EM risk-associated genes, such as *MKNK1* and *TOP3A*, there were other numerous underlying susceptible genes with suggestive evidence for EM that warrant further investigation, as documented in supplementary tables. Third, the different datasets used in the present study showed heterogeneity. To overcome this issue, we applied different statistical methods for multiple corrections of each analysed dataset, such as FDR < 0.05, for pathway enrichment analysis, permuted *P*-value < 0.05, for *Sherlock* analysis, and empirical *P*-value < 0.05, for *in silico* permutation analysis. Moreover, the study participants were only ovarian EM cases, and further studies involving superficial endometriosis and deep EM are needed. Furthermore, to examine whether using the threshold of MAF > 0.0001 affect the results of integrative genomic analyses, we re-performed the Sherlock-based analyses based on SNPs with a MAF > 0.01, and found that these 14 identified EM-risk genes remained to be significant, which is highly consistent with current findings (R2 = 0.9997–0.9999, Supplementary Fig. S8 and Table S11). Finally, our functional experiments on *MKNK1* and *TOP3A* in EECSs are preliminary. Thus, further studies involving animal models and molecular mechanisms are required.

In summary, based on our comprehensive analyses, *MKNK1* and *TOP3A* were identified as EM risk-associated genes, whose genetically modulated abnormal expression may contribute to EM. By combining GWAS summary-based statistics with eQTL-derived regulatory information, this study provides a plausible mechanistic explanation of the functional effects of genetic variants on EM susceptibility. These results provide novel insights into the biological mechanisms of EM and support the promise of translating GWAS findings into new approaches for clinical diagnosis and treatment.

Funding

This work was funded by the National Natural Science Foundation of China (grant Numbers 81901454, 81801420, and 81671426), This work was also funded by Zhejiang Provincial Natural Science Foundation of China (LGF20H040009).

CRediT authorship contribution statement

Yizhou Huang: Conceptualization, Methodology, Software, Data curation, Visualization, Investigation, Validation, Writing – original draft, Writing – review & editing. **Jie Luo:** Conceptualization, Data curation, Investigation, Resources, Writing – original draft. **Yue Zhang:** Methodology, Data curation, Validation, Visualization. **Tao Zhang:** Data curation, Resources, Investigation. **Xiangwei Fei:** Methodology, Funding acquisition. **Liqing Chen:** Data curation, Funding acquisition. **Yingfan Zhu:** Data curation. **Songyue Li:** Data curation. **Caiyun Zhou:** Methodology. **Kaihong Xu:** Funding acquisition. **Yunlong Ma:** Conceptualization, Software, Visualization Writing – review & editing. **Jun Lin:** Supervision, Writing – review & editing. **Jianhong Zhou:** Conceptualization, Supervision, Writing – review & editing.

Data Availability

The GWAS summary data on EM were downloaded from Gene ATLAS (http://geneatlas.roslin.ed.ac.uk/trait/?traits=503), while the eQTL datasets are freely available at http://sherlock.ucsf.edu/submit.html.wm. Further, the RNA-sequencing data underlying this study will be shared on reasonable request to the corresponding author.

Conflict of interest

The authors declare that they have no conflicts of interest.

Acknowledgements

The authors thank all women who provided the sample and clinical information for this study. We would like to thank Editage (www.editage.cn) for English language editing.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.02.001.

References

[1] Zondervan KT, Becker CM, Missmer SA. Endometriosis. (http://). N Engl J Med 2020;382(13):1244–56. https://doi.org/10.1056/NEJMra1810764
[2] Della CL, Di Filippo C, Gabrielli O, Reppuccia S, La Rosa VL, et al. The burden of endometriosis on women's lifespan: a narrative overview on quality of life and psychosocial wellbeing. (http://). Int J Environ Res Public Health 2020;17(13). https://doi.org/10.3390/ijerph17134683
[3] La Rosa VL, De Franciscis P, Barra F, Schiattarella A, Torok P, et al. Quality of life in women with endometriosis: a narrative overview. (http://). Minerva Med 2020;111(1):68–78. https://doi.org/10.23736/S0026-4806.19.06298-0
[4] Saha R, Pettersson HJ, Svedberg P, Olovsson M, Bergqvist A, et al. Heritability of endometriosis. (http://). Fertil Steril 2015;104(4):947–52. https://doi.org/10.1016/j.fertnstert.2015.06.035
[5] Uno S, Zembutsu H, Hirasawa A, Takahashi A, Kubo M, et al. A genome-wide association study identifies genetic variants in the cdkn2bas locus associated with endometriosis in japanese. (http://). Nat Genet 2010;42(8):707–10. https://doi.org/10.1038/ng.612
[6] Nyholt DR, Low SK, Anderson CA, Painter JN, Uno S, et al. Genome-wide association meta-analysis identifies new endometriosis risk loci. (http://). Nat Genet 2012;44(12):1355–9. https://doi.org/10.1038/ng.2445
[7] Painter JN, Anderson CA, Nyholt DR, Macgregor S, Lin J, et al. Genome-wide association study identifies a locus at 7p15.2 associated with endometriosis. (http://). Nat Genet 2011;43(1):51–4. https://doi.org/10.1038/ng.731
[8] Sapkota Y, Steinthorsdottir V, Morris AP, Fassbender A, Rahmioglu N, et al. Meta-analysis identifies five novel loci associated with endometriosis highlighting key genes involved in hormone metabolism. (http://). Nat Commun 2017;8:15539. https://doi.org/10.1038/ncomms15539
[9] Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. (http://). Philos Trans R Soc Lond B Biol Sci 2013;368(1620):20120362. https://doi.org/10.1098/rstb.2012.0362
[10] Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. (http://). Nat Rev Genet 2015;16(4):197–212. https://doi.org/10.1038/nrg3891
[11] Xiang B, Deng C, Qiu F, Li J, Li S, et al. Single cell sequencing analysis identifies genetics-modulated ormdl3(+) cholangiocytes having higher metabolic effects on primary biliary cholangitis. (http://). J Nanobiotechnol. 2021;19(1):406. https://doi.org/10.1186/s12951-021-01154-2
[12] Ma Y, Qiu F, Deng C, Li J, Huang Y, et al. Integrating single-cell sequencing data with gwas summary statistics reveals cd16+monocytes and memory cd8+t cells involved in severe covid-19. (http://). Genome Med 2022;14(1):16. https://doi.org/10.1186/s13073-022-01021-1
[13] Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, et al. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. (http://). Nat Genet 2016;48(5):481–7. https://doi.org/10.1038/ng.3538
[14] Pavlides JM, Zhu Z, Gratten J, Mcrae AF, Wray NR, et al. Predicting gene targets from integrative analyses of summary data from gwas and eqtl studies for 28 human complex traits. (http://). Genome Med 2016;8(1):84. https://doi.org/10.1186/s13073-016-0338-4
[15] Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in uk biobank. (http://). Nat Genet 2018;50(11):1593–9. https://doi.org/10.1038/s41588-018-0248-z
[16] Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The uk10k project identifies rare variants in health and disease. (http://). Nature 2015;526(7571):82–90. https://doi.org/10.1038/nature14962
[17] Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. (http://). Nature 2015;526(7571):68–74. https://doi.org/10.1038/nature15393
[18] Ma X, Wang P, Xu G, Yu F, Ma Y. Integrative genomics analysis of various omics data and networks identify risk genes and variants vulnerable to childhood-onset asthma. (http://). Bmc Med Genom 2020;13(1):123. https://doi.org/10.1186/s12920-020-00768-z
[19] Dong Z, Ma Y, Zhou H, Shi L, Ye G, et al. Integrated genomics analysis highlights important snps and genes implicated in moderate-to-severe asthma based on gwas and eqtl datasets. (http://). Bmc Pulm Med 2020;20(1):270. https://doi.org/10.1186/s12890-020-01303-7
[20] Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. (http://). Am J Hum Genet 2009;85(5):679–91. https://doi.org/10.1016/j.ajhg.2009.09.012
[21] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. (http://). Am J Hum Genet 2007;81(3):559–75. https://doi.org/10.1086/519795
[22] Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, et al. Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. (http://). Plos One 2010;5(5):e10693. https://doi.org/10.1371/journal.pone.0010693
[23] Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. A genome-wide association study of global gene expression. (http://). Nat Genet 2007;39(10):1202–7. https://doi.org/10.1038/ng2109
[24] GTEx Consortium. The genotype-tissue expression (gtex) project. (http://). Nat Genet 2013;45(6):580–5. https://doi.org/10.1038/ng.2653
[25] He X, Fuller CK, Song Y, Meng Q, Zhang B, et al. Sherlock: detecting gene-disease associations by matching patterns of expression qtl and gwas. (http://). Am J Hum Genet 2013;92(5):667–80. https://doi.org/10.1016/j.ajhg.2013.03.022
[26] de Leeuw CA, Mooij JM, Heskes T, Posthuma D. Magma: generalized gene-set analysis of gwas data. (http://). Plos Comput Biol 2015;11(4):e1004219. https://doi.org/10.1371/journal.pcbi.1004219
[27] Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from gwas summary statistics. (http://). Nat Commun 2018;9(1):1825. https://doi.org/10.1038/s41467-018-03621-1
[28] Xie C, Mao X, Huang J, Ding Y, Wu J, et al. Kobas 2.0: a web server for annotation and identification of enriched pathways and diseases. (http://). Nucleic Acids Res 2011;39(Web Server issue):W316–22. https://doi.org/10.1093/nar/gkr483
[29] Af S. Revised american fertility society classification of endometriosis: 1985. (http://). Fertil Steril 1985;43(3):351–2. https://doi.org/10.1016/s0015-0282(16)48430-x

[30] Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. (http://). Nat Biotechnol 2019;37(8):907–15. https://doi.org/10.1038/s41587-019-0201-4
[31] Liao Y, Smyth GK, Shi W. Featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. (http://). Bioinformatics 2014;30(7):923–30. https://doi.org/10.1093/bioinformatics/btt656
[32] Robinson MD, Mccarthy DJ, Smyth GK. Edger: a bioconductor package for differential expression analysis of digital gene expression data. (http://). Bioinformatics 2010;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616
[33] Zhang X, Qi C, Lin J. Enhanced expressions of matrix metalloproteinase (mmp)-2 and -9 and vascular endothelial growth factors (vegf) and increased microvascular density in the endometrial hyperplasia of women with anovulatory dysfunctional uterine bleeding. (http://). Fertil Steril 2010;93(7):2362–7. https://doi.org/10.1016/j.fertnstert.2008.12.142
[34] Remmele W, Stegner HE. [recommendation for uniform definition of an immunoreactive score (irs) for immunohistochemical estrogen receptor detection (er-ica) in breast cancer tissue]. Pathologe 1987;8(3):138–40.
[35] Zhan H, Ma J, Ruan F, Bedaiwy MA, Peng B, et al. Elevated phosphatase of regenerating liver 3 (prl-3) promotes cytoskeleton reorganization, cell migration and invasion in endometrial stromal cells from endometrioma. (http://). Hum Reprod 2016;31(4):723–33. https://doi.org/10.1093/humrep/dew015
[36] Ma Y, Huang Y, Zhao S, Yao Y, Zhang Y, et al. Integrative genomics analysis reveals a 21q22.11 locus contributing risk to covid-19. (http://). Hum Mol Genet 2021;30(13):1247–58. https://doi.org/10.1093/hmg/ddab125
[37] Irungu S, Mavrelos D, Worthington J, Blyuss O, Saridogan E, et al. Discovery of non-invasive biomarkers for the diagnosis of endometriosis. (http://). Clin Proteom 2019;16:14. https://doi.org/10.1186/s12014-019-9235-3
[38] Vestergaard AL, Knudsen UB, Munk T, Rosbach H, Martensen PM. Transcriptional expression of type-i interferon response genes and stability of housekeeping genes in the human endometrium and endometriosis. (http://). Mol Hum Reprod 2011;17(4):243–54. https://doi.org/10.1093/molehr/gaq100
[39] Shchegolev AI, Bykov AG, Faizullina NM, Adamyan LV. Immunohistochemical features of o6-methylguanine-dna methyltransferase expression during ovarian endometriosis. (http://). Bull Exp Biol Med 2018;164(3):386–9. https://doi.org/10.1007/s10517-018-3995-z
[40] Calabrese GM, Mesner LD, Stains JP, Tommasini SM, Horowitz MC, et al. Integrating gwas and co-expression network data identifies bone mineral density genes sptbn1 and mark3 and an osteoblast functional module. (http://). Cell Syst 2017;4(1):46–59. https://doi.org/10.1016/j.cels.2016.10.014
[41] Ma Y, Li MD. Establishment of a strong link between smoking and cancer pathogenesis through dna methylation analysis. (http://). Sci Rep 2017;7(1):1811. https://doi.org/10.1038/s41598-017-01856-4
[42] Xu W, Kannan S, Verma CS, Nacro K. Update on the development of mnk inhibitors as therapeutic agents. (http://). J Med Chem 2022;65(2):983–1007. https://doi.org/10.1021/acs.jmedchem.1c00368
[43] Guo Q, Li VZ, Nichol JN, Huang F, Yang W, et al. Mnk1/nodal signaling promotes invasive progression of breast ductal carcinoma in situ. (http://). Cancer Res 2019;79(7):1646–57. https://doi.org/10.1158/0008-5472.CAN-18-1602
[44] Yang W, Khoury E, Guo Q, Prabhu SA, Emond A, et al. Mnk1 signaling induces an angptl4-mediated gene signature to drive melanoma progression. (http://). Oncogene 2020;39(18):3650–65. https://doi.org/10.1038/s41388-020-1240-5
[45] Ueda T, Watanabe-Fukunaga R, Fukuyama H, Nagata S, Fukunaga R. Mnk2 and mnk1 are essential for constitutive and inducible phosphorylation of eukaryotic initiation factor 4e but not for cell growth or development. (http://). Mol Cell Biol 2004;24(15):6539–49. https://doi.org/10.1128/MCB.24.15.6539-6549.2004
[46] Zhan Y, Guo J, Yang W, Goncalves C, Rzymski T, et al. Mnk1/2 inhibition limits oncogenicity and metastasis of kit-mutant melanoma. (http://). J Clin Invest 2017;127(11):4179–92. https://doi.org/10.1172/JCI91258
[47] Robichaud N, Del RS, Huor B, Alain T, Petruccelli LA, et al. Phosphorylation of eif4e promotes emt and metastasis via translational control of snail and mmp-3. (http://). Oncogene 2015;34(16):2032–42. https://doi.org/10.1038/onc.2014.146
[48] Wang Y, Lyu YL, Wang JC. Dual localization of human dna topoisomerase iiialpha to mitochondria and nucleus. (http://). Proc Natl Acad Sci USA 2002;99(19):12114–9. https://doi.org/10.1073/pnas.192449499
[49] Capranico G, Marinello J, Chillemi G. Type i dna topoisomerases. (http://). J Med Chem 2017;60(6):2169–92. https://doi.org/10.1021/acs.jmedchem.6b00966
[50] Yang J, Bachrati CZ, Ou J, Hickson ID, Brown GW. Human topoisomerase iiialpha is a single-stranded dna decatenase that is stimulated by blm and rmi1. (http://). J Biol Chem 2010;285(28):21426–36. https://doi.org/10.1074/jbc.M110.123216
[51] Nicholls TJ, Nadalutti CA, Motori E, Sommerville EW, Gorman GS, et al. Topoisomerase 3alpha is required for decatenation and segregation of human mtdna. (http://). Mol Cell 2018;69(1):9–23. https://doi.org/10.1016/j.molcel.2017.11.033
[52] Bane K, Desouza J, Shetty D, Choudhary P, Kadam S, et al. Endometrial dna damage response is modulated in endometriosis. (http://). Hum Reprod 2021;36(1):160–74. https://doi.org/10.1093/humrep/deaa255
[53] Kumar A, Dhillon A, Manjegowda MC, Singh N, Mary D, et al. Estrogen suppresses hoxb2 expression via erα in breast cancer cells. (http://). Gene 2021;794:145746. https://doi.org/10.1016/j.gene.2021.145746
[54] Yu HY, Pan SS. Mir-202-5p suppressed cell proliferation, migration and invasion in ovarian cancer via regulating hoxb2. (http://). Eur Rev Med Pharm Sci 2020;24(5):2256–63. https://doi.org/10.26355/eurrev_202003_20491
[55] Liu J, Li S, Cheng X, Du P, Yang Y, et al. Hoxb2 is a putative tumour promotor in human bladder cancer. (http://). Anticancer Res 2019;39(12):6915–21. https://doi.org/10.21873/anticanres.13912
[56] Xu F, Liu Z, Liu R, Lu C, Wang L, et al. Epigenetic induction of tumor stemness via the lipopolysaccharide-tet3-hoxb2 signaling axis in esophageal squamous cell carcinoma. (http://). Cell Commun Signal 2020;18(1):17. https://doi.org/10.1186/s12964-020-0510-8
[57] Luo B, Feng S, Li T, Wang J, Qi Z, et al. Transcription factor hoxb2 upregulates nusap1 to promote the proliferation, invasion and migration of nephroblastoma cells via the pi3k/akt signaling pathway. (http://). Mol Med Rep 2022;25(6). https://doi.org/10.3892/mmr.2022.12721
[58] Lalami I, Abo C, Borghese B, Chapron C, Vaiman D. Genomics of endometriosis: from genome wide association studies to exome sequencing. (http://). Int J Mol Sci 2021;22(14). https://doi.org/10.3390/ijms22147297
[59] Holdsworth-Carson SJ, Fung JN, Luong HT, Sapkota Y, Bowdler LM, et al. Endometrial vezatin and its association with endometriosis risk. (http://). Hum Reprod 2016;31(5):999–1013. https://doi.org/10.1093/humrep/dew047
[60] Powell JE, Fung JN, Shakhbazov K, Sapkota Y, Cloonan N, et al. Endometriosis risk alleles at 1p36.12 act through inverse regulation of cdc42 and linc00339. (http://). Hum Mol Genet 2016;25(22):5046–58. https://doi.org/10.1093/hmg/ddw320
[61] Chou YC, Chen MJ, Chen PH, Chang CW, Yu MH, et al. Integration of genome-wide association study and expression quantitative trait locus mapping for identification of endometriosis-associated genes. (http://). Sci Rep 2021;11(1):478. https://doi.org/10.1038/s41598-020-79515-4
[62] Yang CP, Li X, Wu Y, Shen Q, Zeng Y, et al. Comprehensive integrative analyses identify glt8d1 and csnk2b as schizophrenia risk genes. (http://). Nat Commun 2018;9(1):838. https://doi.org/10.1038/s41467-018-03247-3
[63] Zhong J, Li S, Zeng W, Li X, Gu C, et al. Integration of gwas and brain eqtl identifies flot1 as a risk gene for major depressive disorder. (http://). Neuropsychopharmacology 2019;44(9):1542–51. https://doi.org/10.1038/s41386-019-0345-4