

RESEARCH

Open Access



Rare variant analyses provide insights into the genetic architecture of endometriosis

Yuanyuan Gui^{1,2†}, Yuhan Xie^{3†}, Xiaomei Tong^{1,2*} and Hongyu Zhao^{3*}

Abstract

Background Endometriosis (EMs) is a prevalent gynecologic disorder that often causes severe chronic pelvic pain and infertility. Although genome-wide association studies have identified common variants of EMs, the contribution of rare variants remains unclear.

Results We conducted an exome-wide analysis in 87,100 unrelated female participants of European ancestry from the UK Biobank. Single variant association tests for rare variants and gene-based analyses were performed, with replication in the FinnGen and Biobank Japan cohorts. No single rare variants achieved genome-wide significance ($p < 5 \times 10^{-8}$). While gene-based analyses identified *SOGA1* significantly associated with EMs at the Bonferroni-corrected threshold of $p < 2.98 \times 10^{-6}$, which was replicated in both external cohorts. Functional analyses indicated involvement of implicated genes and interaction networks in regulating gluconeogenesis, muscle filament sliding, and autophagosome. We also observed expression differences of *SOGA1* across various cell types and found significant cell type-specific enrichments for EMs.

Conclusions This study establishes the contribution of rare genetic variants to endometriosis pathogenesis through the discovery of a novel susceptibility gene, advancing the genetic understanding of this disease.

Keywords Genomics, Endometriosis, Whole exome sequencing, *SOGA1*

Background

Endometriosis (EMs) is a chronic, inflammatory condition characterized by the presence of endometrial-like tissue outside the uterus, affecting approximately 10% of reproductive-age women globally. This disease often results in severe pelvic pain, infertility, and reduced quality of life, with diagnostic delays of 5–12 years further exacerbating patient burden [1]. The etiology of EMs remains unclear. While several hypotheses have been proposed, none fully explains the disease's development across all observed sites.

Endometriosis is estimated to have a heritability of 47–51% based on twin studies in Swedish cohorts [2]. Danish population studies show daughters of women with endometriosis are more than twice as likely to develop the condition [3], indicating a strong genetic

[†]Yuanyuan Gui and Yuhan Xie have contributed equally to this work.

*Correspondence:

Xiaomei Tong
3406028@zju.edu.cn
Hongyu Zhao
hongyu.zhao@yale.edu

¹Assisted Reproduction Unit, Department of Obstetrics and Gynecology, School of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, 3 East Qingchun Road, Hangzhou 310016, Zhejiang, China

²Present address: Zhejiang Provincial Clinical Research Center for Reproductive Health and Disease, Hangzhou, China

³Present address: Department of Biostatistics, Yale School of Public Health, Yale University, 60 College Street, New Haven, CT 06520, USA



© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

component in its pathogenesis. Genome-wide association studies (GWAS) have identified 42 significant risk loci (49 independent signals) through meta-analyses of 60,674 European and East Asian cases and 701,926 controls, with the strongest effects in the American Society of Reproductive Medicine (ASRM) stage III/IV disease, yet these explain only 5.01% of disease variance [4]. Koller et al. [5] identified one locus in the UK Biobank data from female participants of European descent. Kloeve-Mogensen et al. [6] calculated a polygenic risk score (PRS) using 14 single-nucleotide polymorphisms (SNPs) from a GWAS meta-analysis of 17,000 EM European cases, achieving modest prediction ($AUC=0.64$). The focus of the genetic architecture of endometriosis has mainly been on common variants. While the common variants collectively account for ~26% of heritability [7], a substantial proportion of genetic risk remains unexplained.

The substantial missing heritability in complex diseases has motivated increasing interest in rare, functional variants [8]. Some reports indicate that rare variants (especially large-effect functional loss or destructive missense mutations) can directly identify key pathogenic pathways and core genes that are often obscured in common variant studies due to extreme polygenicity. Whereas others suggest their overall contribution may be modest [9, 10]. Despite these differences, rare loss-of-function and damaging missense variants remain powerful tools for identifying core pathogenic mechanisms, and convergence between rare and common variant signals has been observed across multiple complex traits. These considerations highlight the value of investigating rare coding variation in endometriosis.

In this study, we investigated the contribution of rare variants to EMs and uncovered potential pathogenic mechanisms. Utilizing exome sequence data from the UK Biobank (UKB), we identified genetic associations with EMs through single-variant and gene-based analyses. Our findings were replicated in individuals of Finnish and Japanese ancestry. Additionally, we explored the functional characterization of the identified gene and examined its expression using single-cell RNA expression datasets of EMs.

Methods

Study cohorts

The UKB [11] is a large-scale study that recruited participants across the United Kingdom and was approved by the North West Multi-Centre Research Ethics Committee as a Research Tissue Bank, and all participants provided informed consent. Whole exome sequencing (WES) data, array genotypes data and phenotype data used in our study were accessed under application number 29900. Genome-wide association statistics of

FinnGen [12] and Biobank Japan (BBJ) [13] are previously published.

Genomic data

For this study, we included ~200,000 participants with available array genotype, WES, and phenotype data. We identified 3,736 individuals with endometriosis based on the ICD-10 codes N80 (Field 41,270, 41,280) and self-reported diagnoses of 1402 (Field 20,002). 196,907 individuals with no N80 diagnosis and no self-reported endometriosis were treated as controls. Only females of European ancestry were included (Field 22,006, 21,000). Subjects were excluded if there was a mismatch between genetic sex (Field 22,001) and self-reported sex (Field 31), with extreme heterozygosity (Field 22,027) or close relatives (kinship coefficient >0.0442). This retained 87,100 individuals for subsequent analyses.

Association analyses

Genetic association analyses were performed using REG-ENIE [14], a fast two-step whole-genome regression method that effectively handles unbalanced case-control phenotypes, population structure and relatedness. We included in step1 of REGENIE array variants with minor allele frequency ($MAF \geq 1\%$), $\geq 90\%$ genotyping rate, Hardy-Weinberg equilibrium test $P \geq 10^{-15}$, minor allele count ($MAC \geq 100$), not in long-range linkage disequilibrium (LD) regions, and LD pruning (1,000 variant windows, 100 variant sliding windows and $r^2 < 0.9$). These thresholds follow the UKB analysis recommendations in the REGENIE documentation and the exome sequencing analysis by Backman et al. [10]. We adjusted for age at recruitment and the first four genetic principal components (PCs) to account for population structure. These PCs were calculated using PLINK v2.0 [15] from the quality-controlled array data. In step 2, approximate Firth logistic regression was applied to test rare exome variants ($MAF < 1\%$), and we included additionally four PCs derived from rare variants as previous research showed that PCs from common variants inadequately adjust for fine-scale population structure [16, 17]. The Genetic variants with P values $< 5 \times 10^{-8}$ were considered significant.

Gene-based association analysis was performed using the burden test in step 2 of REGENIE to aggregate the effects of rare variants within genes, thereby increasing statistical power for detecting associations driven by rare functional variation. We constructed a mask consisting of rare ($MAF < 1\%$) loss-of-function (LoF) variants and deleterious missense (Dmis) variants, defined following Xie et al. [18], based on concordant predictions from nine in silico tools, including SIFT [19], PolyPhen-2 (HVAR and HDIV), M-CAP [20, 21], MetaLR, MetaSVM [22], LRT [23], PROVEAN, and MutationTaster [24, 25]. The

covariates were the same as those used in step 2 of the exome-wide single-variant tests. The significance level was set as 0.05 divided by the number of genes (16,773) examined.

Validations

To evaluate the robustness and generalizability of the gene-based findings across different populations, we downloaded two independent publicly available GWAS summary statistics from FinnGen and BBJ of endometriosis, with cases defined by ICD codes in FinnGen and by clinical diagnosis in the BBJ. As individual-level genotype data were not accessible, we performed gene-based association testing using GATES [26], which combines SNP-level association evidence across genes while accounting for LD. The 1000 Genomes Project phase 3 FIN and JPT population datasets were used as LD reference panels to ensure appropriate adjustment for population-specific genetic variation. Other parameters adhered to the default settings of GATES.

Pathway analysis

Pathway analyses were performed using the clusterProfiler R package [27] to identify enriched biological processes of relevant genes. Gene Ontology (GO) [28] terms were analyzed. For these analyses, we selected the top genes ($P < 5 \times 10^{-5}$) based on gene-based associations, as well as genes retrieved from BioGRID [29] that interact with the identified candidate genes.

PheWAS analysis

To explore the broader impact of the candidate risk genes on health outcomes, we conducted a phenome-wide association study (PheWAS). The ICD-10 data from the UK Biobank were downloaded using data field ID 41720 and preprocessed as a phenotype table using the PheWAS package [30]. ICD codes were mapped to phecodes using mapping tables [31], available at <https://wei-lab.app.vumc.org/phecode>. In total, 1532 phecodes were mapped. Carrier status for each identified candidate gene was determined as True if the participant had at least one LoF or Dmis variant in the exonic region of the gene, and as False otherwise. Age, sex, and the top four PCs were adjusted in the PheWAS analysis to account for demographic and population structure.

Single-cell transcriptomic analyses

To identify EMs relevant cell types and evaluate expression patterns of candidate genes, we performed cell type specific [32] and differential expression analyses using publicly available single-cell RNA sequencing (scRNA-seq) datasets from Fonseca et al. [33] and Tan Y et al. [34]. For the cell type-specific analysis, we examined nine annotated cell populations (endothelial, epithelial,

smooth muscle, mesenchymal, myeloid, mast, B/plasma, erythrocytes and T/NK cells) across five lesions in Fonseca et al. dataset, and five major populations (stromal, endothelial, lymphocytes, epithelial and myeloid) from four lesions in Tan Y et al. dataset, respectively. To further assess the consistency between the two datasets, the nine cell types from Fonseca et al. were grouped into six broader categories (stromal, endothelial, lymphocytes, epithelial, myeloid and erythrocytes), corresponding to those reported by Tan et al. Differential expression analyses were subsequently performed using the Wilcoxon test. Details of the single-cell transcriptome acquisition and processing are reported in the original studies. A workflow for the data analyses is presented in Fig. 1.

Results

Genetic association analyses

The basic characteristics of participants in this study are shown in Table 1.

We performed single variants analyses for the 1,423,086 rare variants with MAC above 8. Genomic inflation was 1.05. There was no statistically significant association between tested variants and EMs at a genome-wide significance level ($p < 5 \times 10^{-8}$). Only one variant in LMO1 was marginally significant ($p = 4.84 \times 10^{-7}$, Fig. S1, Table S1).

We aggregated 16,778,137 LoF and Dmis variants into gene-based tests using REGENIE, corresponding to 16,773 genes. One gene (*SOGA1*) was significantly associated with EMs after Bonferroni correction ($p = 1.32 \times 10^{-6}$, $< 0.05/16,773$, Table S2). There were 15 LoF rare variants in total distributed in *SOGA1* in this test (Table S3). Two additional genes, MYBPC2 ($p = 1.53 \times 10^{-5}$) and RAB19 ($p = 4.51 \times 10^{-5}$) reached suggestive significance (Table S2). These results were presented using a quantile–quantile (Q-Q) plot and a Manhattan plot (Fig. 2A and B).

Replication Analyses

We replicated the association between *SOGA1* and EMs risk using gene-based testing in FinnGen and BBJ, respectively. The *SOGA1* was associated with EMs ($p = 0.015$ and $p = 0.011$, respectively, in Finnish and Japanese populations) at a nominal significance level of 0.05 (Table S4).

Interaction Networks Analysis

We searched BioGRID 4.4 for interactions involving *SOGA1*. A total of 130 interactors with 165 interactions were constructed (Fig. S2). All interactions were curated from published literature-based evidence. After mapping these interactors to the FinnGen and BBJ GWAS summary statistics, 116 genes were included in the gene-based analyses. Using GATES, 11 of the 116 genes

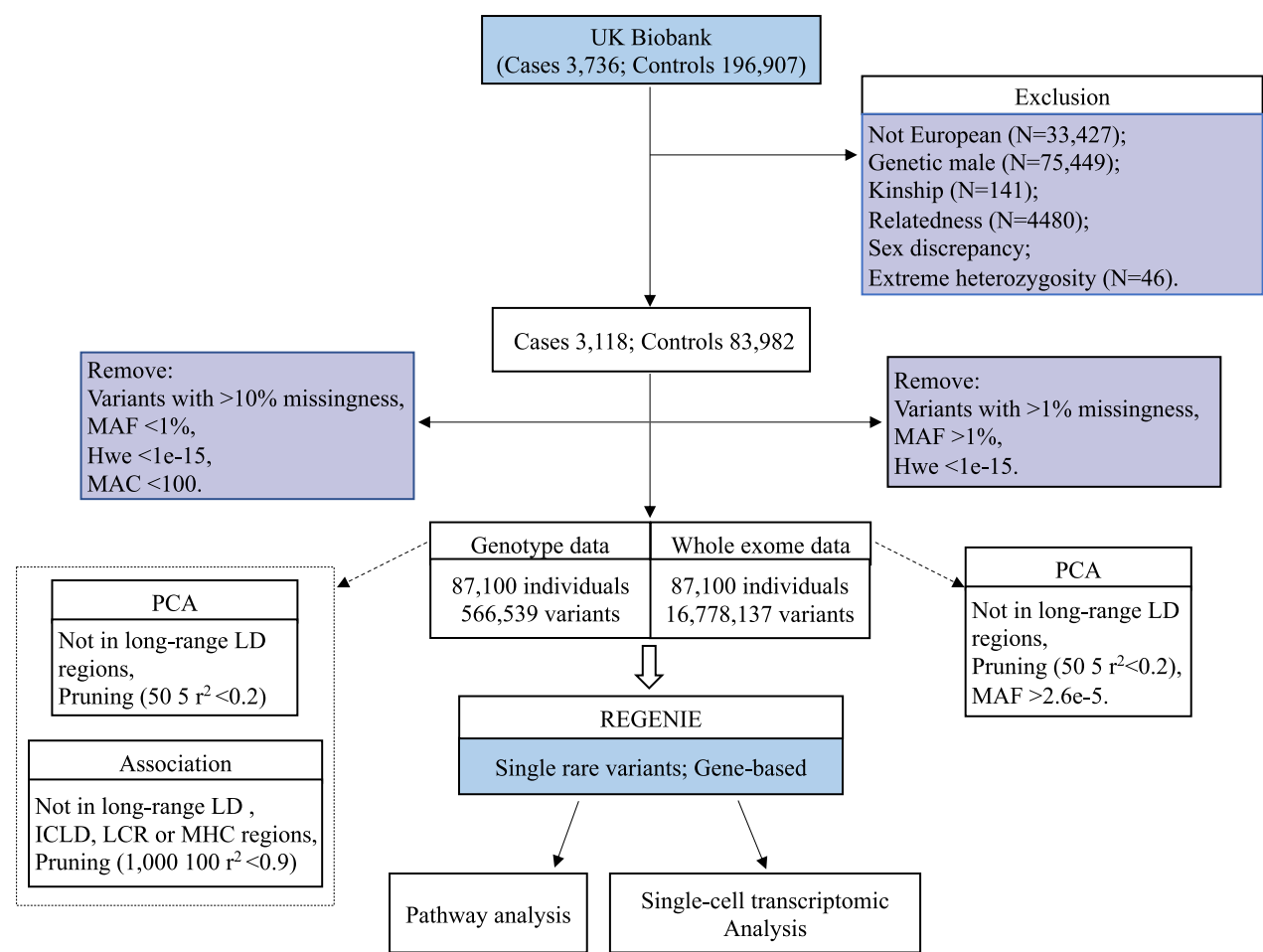


Fig. 1 A workflow for the data analyses in this study

Table 1 Participant characteristics

	Sample size	Age, y; mean (SD)		Sex	An- cestry
	(Cases/controls)	Cases	Controls		
Discovery cohort					
UK Biobank	3,118/83,982	53.5(8.0)	56.7(8.0)	Fe- males	Euro- pean
Replication cohort					
Finn-Gen	10,029/81,593	–	–	Fe- males	Finnish
Bio-bank Japan	734/102,372	42.5 (8.1)	61.4 (14.3)	Fe- males	Japa- nese

showed significant associations with EMs at the threshold of $p < 0.05$. This corresponds to a two-fold enrichment (11/5.8) with a statistical significance of 0.031. Several of these genes, including *GSK3B* [35], *PPP2CA* [36], *ESR2* [37], *SAMHD1* [38], *CYLD* [39], and *AURKA* [40], have also been reported to be associated with endometriosis in previous studies. Of the 11 genes, six were significant in the FinnGen analysis, and five were significant in the BBJ analysis (Table S4), showing very high enrichments

(6/0.55 with p-value of 6×10^{-7} for FinnGen and 5/0.55 with p-value of 1.1×10^{-4} for BBJ).

Pathway Analysis

Three genes in the gene-based analyses that were significantly or suggestively associated with EMs were examined in GO terms. In GO term enrichment analysis, the identified significant terms included gluconeogenesis, relevant biological processes, muscle filament sliding, and autophagosome assembly ($p_{adjust} = 3.29 \times 10^{-3} \sim 4.4 \times 10^{-2}$). We also looked up *SOGA1* in previously reported literature. As described [41, 42], *SOGA1* was identified as a microtubule-associated factor, related to chromosome segregation. It is also involved in autophagy and glucose.

We further examined the GO terms of *SOGA1* related gene sets identified in the interaction networks. Ten genes related to significant GO terms and 21 significant biological processes/cellular components/molecular functions that might be involved in the migration and cell proliferation of EMs. Most of the enriched GO terms were related to Regulation of Protein Binding

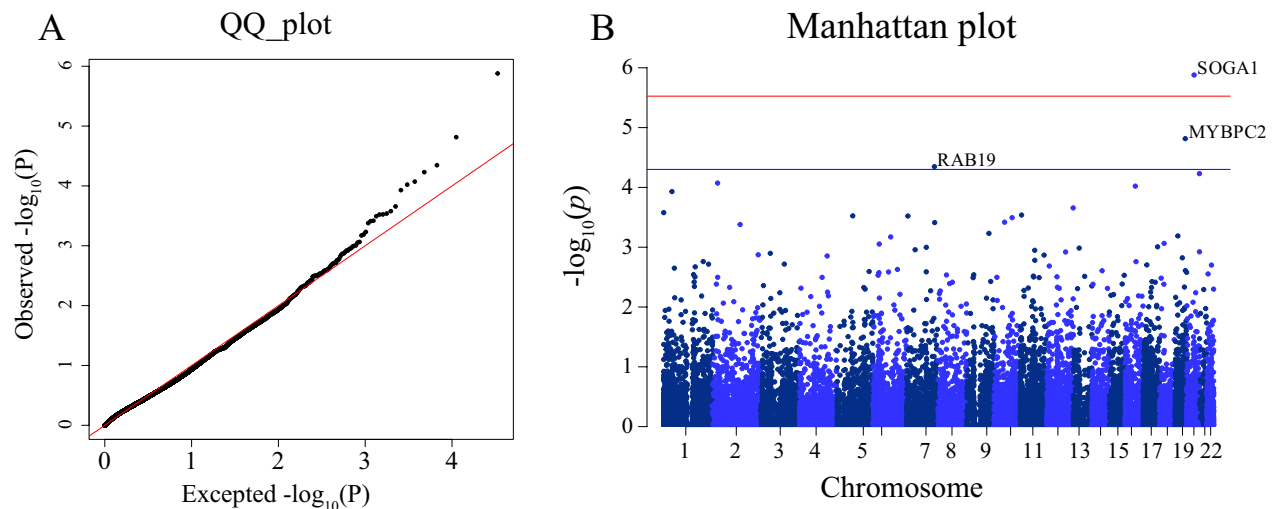


Fig. 2 QQ plot and Manhattan plot for exome-wide gene-based analysis. **A** Q-Q plot shows observed vs. expected p-values in exome-wide association analysis. **B** Manhattan plots show the genome along the x-axis and the gene-based association significance levels on the y-axis. Red line represents the Bonferroni corrected threshold with $P = 2.98 \times 10^{-6}$, blue line represents the suggestive significance level at $P = 5 \times 10^{-5}$

($p_{\text{adjust}} = 6.21 \times 10^{-3}$), Wnt Signaling/Beta-Catenin Destruction ($p_{\text{adjust}} = 4.42 \times 10^{-2}$), Cortical Actin Cytoskeleton ($p_{\text{adjust}} = 3.03 \times 10^{-2}$), Spindle/Spindle Pole/Mitotic Spindle Midzone/Spindle Pole Centrosome ($p_{\text{adjust}} = 4.42 \times 10^{-2}$).

Phenome-wide association study

We tested *SOGA1* identified in gene-based analysis for association with 1,532 traits. *SOGA1* was statistically significantly associated with Endometriosis ($p = 2.47 \times 10^{-6}$, $<0.05/1532$), Dermatophytosis of the body ($p = 3.31 \times 10^{-6}$) and Noninflammatory disorders of cervix ($p = 2.50 \times 10^{-5}$) (Fig. 3).

Single-cell transcriptomic Analysis

We identified EMs relevant cell types by analyzing scRNA-seq expression datasets together with summary statistics of EMs. For scRNA-seq from Fonseca et al., EMs have been categorized into five major classes of tissue types, and the corresponding cell types in these datasets have been classified into nine categories. Of the five endometrial types, endometrioma exhibited epithelial and mesenchymal cell-type enrichments ($p = 0.01$, $p = 0.02$), no endometriosis detected exhibited epithelial-specific enrichments ($p = 0.03$), and the unaffected ovary exhibited strong enrichments in endothelial cells ($p = 0.03$). When the five endometrial types were considered together, no statistically significant enrichments were observed for any specific cell type ($p = 0.06 \sim 0.99$, Fig. 4A). The nine identified cell types were then grouped into six broader categories, including stromal, endothelial, lymphocytes, epithelial, myeloid and erythrocytes. We found significant enrichments in stromal, epithelial and endothelial relative to

immune and erythrocytes for most endometrial types ($p = 0.01 \sim 0.03$, Fig. 4A). Additionally, there were three marginally significant enrichments in stromal cells ($p = 0.05$). For the scRNA-seq dataset from Tan Y et al., stromal enrichment was also observed (Fig. 4B).

We further evaluated *SOGA1* gene expression using these datasets. For nine cell types, *SOGA1* was mostly highly expressed in endothelial cells and showed relatively medium expressions in smooth muscle, mesenchymal and epithelial cells in the context of five endometrial-types and combined EMs group (Fig. 4C). While their $\log_2\text{FC}$ values differed by less than 0.29. For six cell types in Fonseca et al. as well as five cell types in Tan Y et al., there was a higher expression of *SOGA1* in stromal, endothelial and epithelial cells (Fig. 4D and E).

Discussion

Our WES study identified a gene-wide significant association with *SOGA1* in a gene-based test including 15 LoF rare variants ($p = 1.32 \times 10^{-6}$). Consistent results were also obtained using SKAT and SKAT-O methods implemented in REGENIE, which allow for heterogeneous effect directions (data not shown). The association was supported at nominal significance in the Finnish and East Asian cohorts ($p = 0.015$ and $p = 0.011$). As expected, we did not observe associations with single variants in this study due to the sample size limitation.

We applied GATES to FinnGen and BBJ GWAS summary statistics to test for gene-based associations between *SOGA1* interacting genes and endometriosis. Eleven of the 116 evaluable genes showed significant associations, representing a two-fold enrichment. Several of these genes, including *GSK3B*, *PPP2CA*, *ESR2*, *SAMHD1*, *CYLD*, and *AURKA*, have also been linked

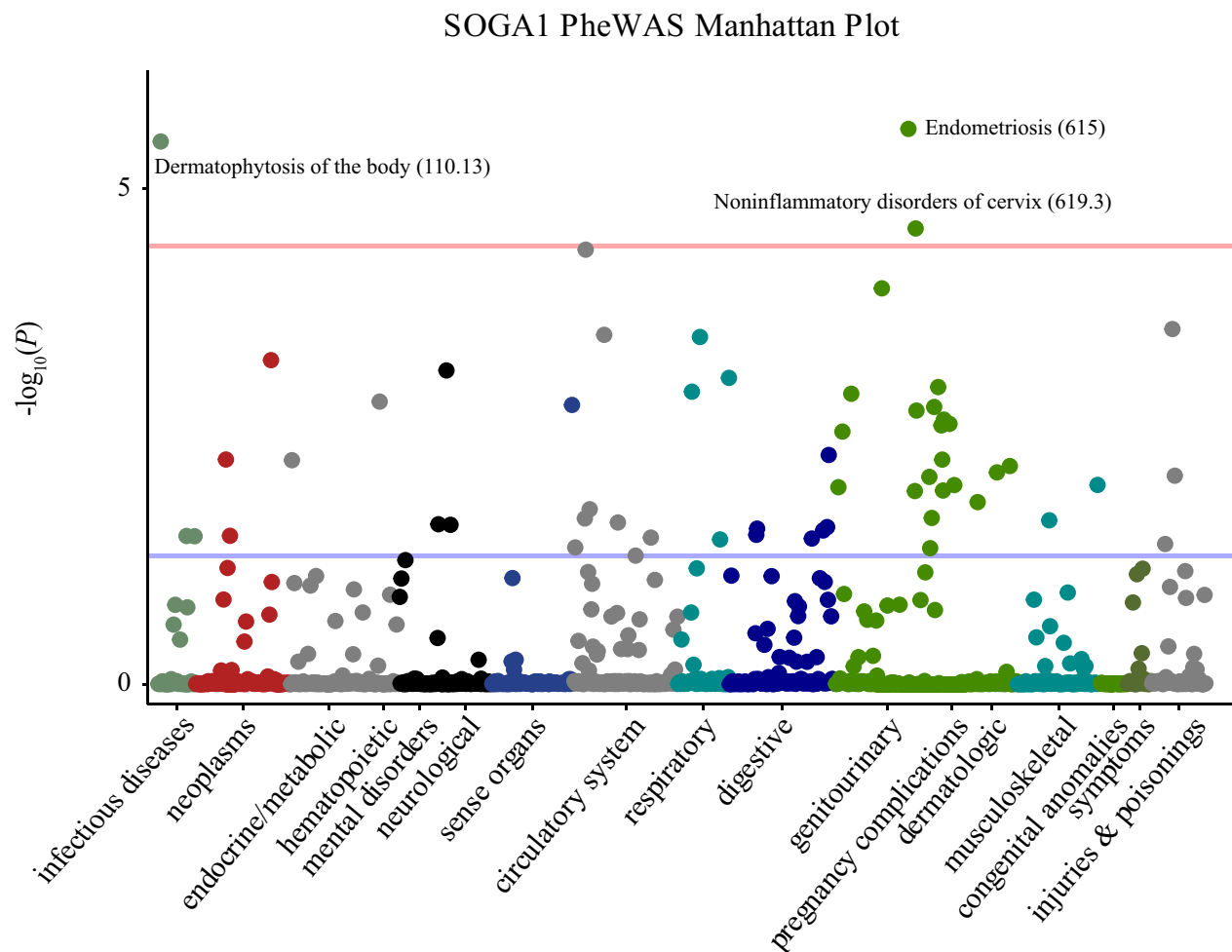


Fig. 3 Manhattan Plot of PheWAS analysis. This plot showed the relationship between *SOGA1* and tested phenotypes. The x-axis represents the phenotypes tested, color-coded by ICD-10 categories, while the y-axis shows the $-\log_{10}(p\text{-values})$ of the association significance. The red line represents the significance threshold (0.05/1,532), and the blue line represents the suggestive significance threshold (0.05)

to endometriosis in previous studies, providing independent support for their relevance. Functional enrichment analyses of the *SOGA1*-interacting gene set suggested potential involvement in pathways related to glucose metabolism and autophagy, although these signals remain exploratory and should not be interpreted as direct mechanistic evidence. Previous experimental studies have shown that *SOGA1* encodes a microtubule-associated protein that integrates centrosomal and Golgi microtubules on the Golgi membrane, aiding in directional cell migration [41, 43]. *SOGA1* is essential for proper chromosome segregation during mitosis and has been demonstrated to regulate autophagy and glucose homeostasis by suppressing glucose production in an adiponectin- and insulin-dependent manner [44, 45]. These biological functions overlap with several well-recognized pathophysiological processes in endometriosis, including autophagy, fibrosis, angiogenesis, migration, and glucose metabolism [46, 47]. A recent GWAS meta-analysis

identified 42 significant risk loci associated with EMs [4]. Although our identified gene does not overlap with those of large GWAS studies focused on common variants, it can be partly attributed to our focus on rare variants ($MAF < 0.01$), which may not have been well covered by GWAS arrays or imputation. These studies also emphasized the diverse biological pathways implicated in endometriosis.

Endometriosis is characterized by abnormal migration and invasion of endometrial-like tissue outside the uterus, which involves complex cell migration mechanisms. To explore potential cellular contexts in which *SOGA1* may act, we examined its expression across cell types using publicly available scRNA-seq datasets. *SOGA1* was observed highly expressed in endothelial and stromal cells in EMs lesions, and LDSC-SEG analyses identified these cell types were relevant to EMs. Endothelial and stromal cells are recognized contributors to the migratory, invasive, and tissue-remodeling behavior

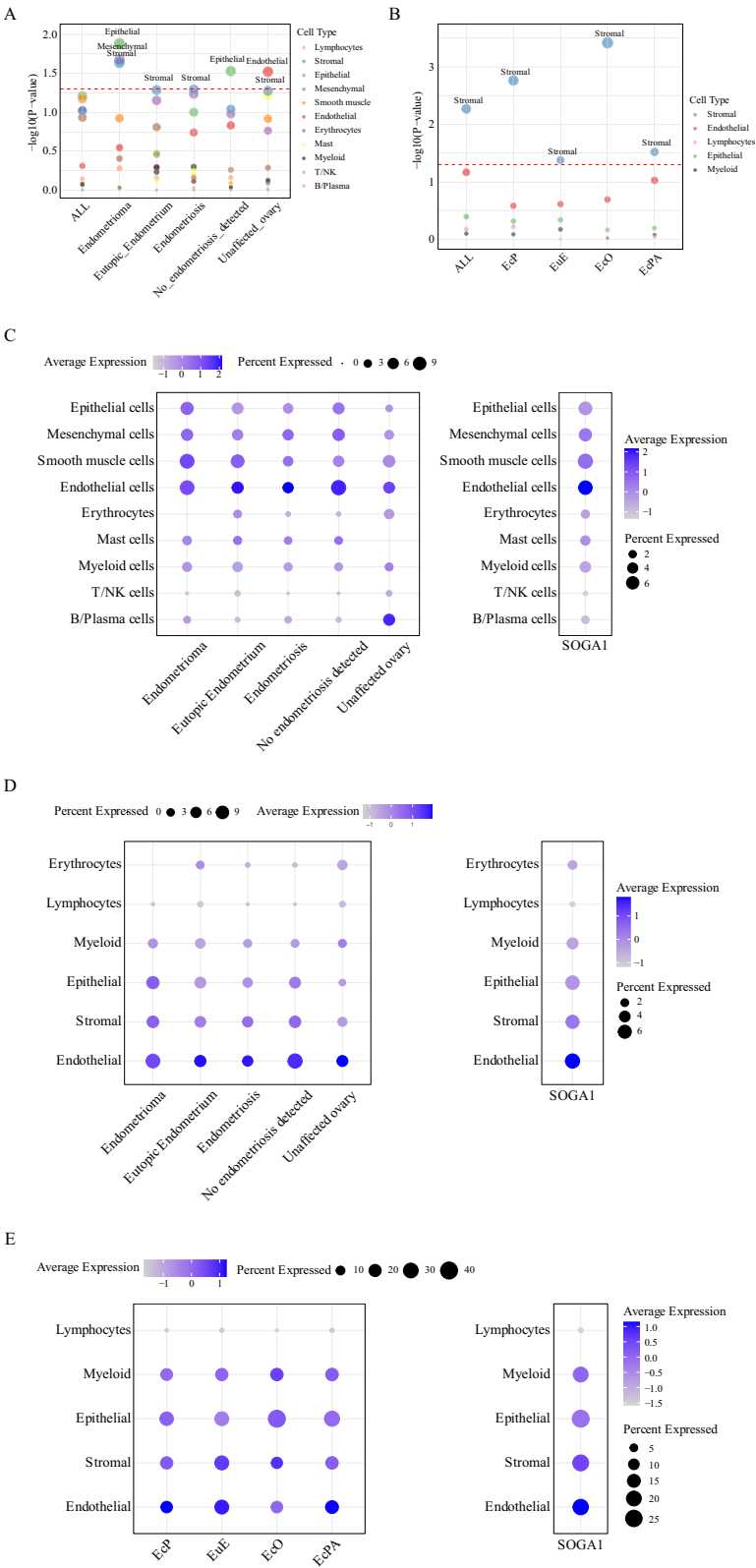


Fig. 4 Heritability enrichment of endometriosis and differential expression of *SOGA1* across cell types. **A** Heritability enrichment of endometriosis in nine cell types and six broader cell groups evaluated by LDSC. The color of dots represents different cell groups. **B** Heritability enrichment of endometriosis in five cell groups based on data from Tan Y et al. **C–E** Dot plots showing *SOGA1* expression levels and the proportion of expressing cells across various cell types. Dot size indicates the proportion of *SOGA1*-expressing cells, and color intensity reflects expression levels. The left side displays lesion-specific groups, and the right side shows the combined EMs group

observed in EMs [48–51]. Our data do not provide direct evidence that *SOGAI* mediates these processes. Nevertheless, the enrichment of *SOGAI* expression in these cells raises the possibility that it may contribute to behaviors involved in lesion development, such as migration, invasion, or tissue remodeling. Determining whether *SOGAI* has a mechanistic role in these pathways will require further functional investigation.

Our findings suggest that rare variants in *SOGAI* may contribute to EMs pathogenesis via altered glucose metabolism, autophagy, and cell migration. Although PheWAS results showed pleiotropic effects of *SOGAI* on noninflammatory disorders of cervix, the consistently low *SOGAI* expression in immune-related cells across most lesions of EMs, suggest that *SOGAI* may not exert its effects on endometriosis through direct immune pathways. Clinically, identifying *SOGAI*-related pathways could guide therapeutic strategies targeting metabolic and migratory processes in EMs. However, additional biological validation is required before translation to practice.

The main strength of this study lies in its large WES dataset (87,100 participants), enabling rare variant analysis at an unprecedented scale in EMs. Replication in independent cohorts of different ancestries adds robustness. Integration with scRNA-seq data further strengthens the biological plausibility of *SOGAI*.

Nevertheless, several limitations of this study should be acknowledged. First, while our findings strongly implicate *SOGAI*, definitive functional validation remains essential to establish causal links to disease processes. Our team is actively collecting intraoperative tissues (ectopic/eutopic endometrium), follicular fluid, granulosa cells, and peripheral blood, to enable future functional and molecular investigations.

Second, endometriosis is a clinically heterogeneous disorder comprising multiple subtypes with both shared and distinct biological mechanisms. In population-based resources such as the UKB, phenotype misclassification represents a major challenge. Asymptomatic or minimally symptomatic endometriosis is not uncommon, leading to contamination of controls [52]. In addition, phenotype definitions in population-based cohorts often rely on clinical assessment or self-report rather than surgical and histological confirmation. Even when laparoscopy is performed, visual inspection alone has limited sensitivity and specificity, with well-documented false-positive and false-negative rates [53, 54], further complicating reliable phenotype assignment. Consistent with these limitations, the prevalence of endometriosis observed in this study (3.6%) is substantially lower than epidemiological estimates of approximately 10%, suggesting considerable underdiagnosis, particularly among controls. Such misclassification is expected to bias effect

estimates toward the null and reduce statistical power. These challenges are especially consequential for rare-variant association analyses. Individual rare variants within *SOGAI* were observed at extremely low frequencies, typically one or two heterozygous carriers across the entire cohort. Under this allelic architecture, even modest phenotype misclassification or sample restriction can lead to a disproportionate loss of statistical power. Narrowing the phenotype may increase heritability and strengthen genetic signals, but formal subtype-specific analyses were not feasible in the present study due to limited phenotypic annotation and sample size. Instead, we performed a series of restriction-based sensitivity analyses, including exclusion of overlapping diagnoses (N70, N73, K58, and N30.1), restriction of cases to those diagnosed before 35 years of age, and more conservative control definitions excluding individuals with any ICD-10 diagnosis records. While these restrictions improved phenotypic specificity, they also resulted in pronounced reductions in effective sample size, particularly among cases. Under such conditions, statistical power for rare-variant burden tests collapses nonlinearly, and association results become highly sensitive to stochastic variation in carrier distribution. Consequently, increasingly stringent phenotype definitions did not lead to systematic strengthening of association signals but instead produced instability across single-variant and gene-based results. This pattern reflects a fundamental trade-off between phenotype refinement and statistical power in the context of allelic rarity.

Finally, although replication analyses were performed in independent cohorts, the results should be interpreted with caution. Due to the lack of independent rare-variant summary statistics for rare-variant association replication, gene-based testing relied on common-variant summary statistics aggregated using GATES. As such, this analysis does not represent a direct replication of rare-variant effects but rather provides supportive gene-level evidence driven by common variants. Future replication efforts using independent sequencing-based cohorts will be critical for confirming rare-variant associations.

Conclusions

In our study, we used WES data from 87,100 participants in the UK Biobank, identifying a significant association between *SOGAI* and the risk of EMs in European ancestry populations. The association was also repeated in both Finnish and Japanese cohorts. Our findings will provide insight into disease mechanisms and targets for biological experiments to gain further understanding about the role of *SOGAI* in EMs pathogenesis.

Abbreviations

EMs	Endometriosis
UKB	UK Biobank

WES	Whole exome sequencing
LoF	Loss-of-function
Dmis	Deleterious missense
PheWAS	Phenome-wide association study
scRNA-seq	Single-cell RNA sequencing

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40246-026-00934-x>.

Supplementary material 1.

Supplementary material 2.

Acknowledgements

The UK Biobank Resource used in this article was obtained from Application Number 29900.

Author contributions

Yuanyuan Gui, Yuhua Xie: Conceptualization, Methodology. Xiaomei: Administrative and technical support services. Hongyu Zhao: Supervision.

Funding

Not applicable.

Availability of data and materials

The data used in this study are publicly available. Application to UK Biobank grants access to individual-level data. Summary statistics for endometriosis were obtained from FinnGen (<https://www.finnngen.fi/en>, version 6, N14_ENDOMETRIOSIS) and Biobank Japan (<http://jenger.riken.jp/en/result>). Endometriosis single-cell RNA-sequencing data were obtained from <https://github.com/lawrenson-lab/AtlasEndometriosis> and GEO accession GSE179640. All resources were accessed between June 2024 and June 2025, and analyses were conducted using the versions available at the time of access.

Declarations

Ethics approval and consent to participate

The UK Biobank study has approval from the North West Multi-Centre Research Ethics Committee as a Research Tissue Bank approval, and all participants provided informed consent. This study was conducted in accordance with the ethical standards of local institutional review board and with the 1964 Helsinki Declaration and its later amendments.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 17 December 2025 / Accepted: 10 February 2026

Published online: 24 February 2026

References

- As-Sanie S, Mackenzie SC, Morrison L, et al. Endometriosis: a review. *JAMA*. 2025;334(1):64–78.
- Saha R, Pettersson HJ, Svedberg P, et al. Heritability of endometriosis. *Fertil Steril*. 2015;104(4):947–52.
- Dalsgaard T, Hjordt Hansen MV, Hartwell D, Lidegaard O. Reproductive prognosis in daughters of women with and without endometriosis. *Hum Reprod*. 2013;28(8):2284–8.
- Rahmioglu N, Mortlock S, Ghiasi M, et al. The genetic basis of endometriosis and comorbidity with other pain and inflammatory conditions. *Nat Genet*. 2023;55(3):423–36.
- Koller D, Pathak GA, Wendt FR, et al. Epidemiologic and genetic associations of endometriosis with depression, anxiety, and eating disorders. *JAMA Netw Open*. 2023;6(1):e2251214.
- Kloeve-Mogensen K, Rohde PD, Twisttman S, et al. Polygenic Risk Score Prediction for Endometriosis. *Front Reprod Health*. 2021;3:793226.
- Lee SH, Harold D, Nyholt DR, et al. Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. *Hum Mol Genet*. 2013;22(4):832–41.
- Pathan N, Deng WQ, Di Scipio M, et al. A method to estimate the contribution of rare coding variants to complex trait heritability. *Nat Commun*. 2024;15(1):1245.
- Weiner DJ, Nadig A, Jagadeesh KA, et al. Polygenic architecture of rare coding variation across 394,783 exomes. *Nature*. 2023;614(7948):492–9.
- Backman JD, Li AH, Marcketta A, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*. 2021;599(7886):628–34.
- Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779.
- Kurki MI, Karjalainen J, Palta P, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*. 2023;613(7944):508–18.
- Ishigaki K, Akiyama M, Kanai M, et al. Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat Genet*. 2020;52(7):669–79.
- Mbatchou J, Barnard L, Backman J, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet*. 2021;53(7):1097–103.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4:7.
- Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet*. 2012;44(3):243–6.
- Zaidi AA, Mathieson I. Demographic history mediates the effect of stratification on polygenic scores. *Elife*. 2020. <https://doi.org/10.7554/eLife.61548>.
- Xie Y, Acosta JN, Ye Y, et al. Whole-exome sequencing analyses support a role of vitamin D metabolism in ischemic stroke. *Stroke*. 2023;54(3):800–9.
- Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. *Nat Protoc*. 2016;11(1):1–9.
- Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7(4):248–9.
- Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet*. 2016;48(12):1581–6.
- Dong C, Wei P, Jian X, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet*. 2015;24(8):2125–37.
- Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res*. 2009;19(9):1553–61.
- Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE*. 2012;7(10):e46688.
- Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods*. 2010;7(8):575–6.
- Li MX, Gui HS, Kwan JS, Sham PC. GATES: a rapid and powerful gene-based association test using extended Simes procedure. *Am J Hum Genet*. 2011;88(3):283–93.
- Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16(5):284–7.
- Gene Ontology C, Aleksander SA, Balhoff J, et al. The gene ontology knowledgebase in 2023. *Genetics*. 2023;224(1).
- Oughtred R, Stark C, Breitkreutz BJ, et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res*. 2019;47(D1):D529–41.
- Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*. 2014;30(16):2375–6.
- Wu P, Gifford A, Meng X, et al. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: workflow development and initial evaluation. *J Med Internet Res*. 2019;7(4):e14325.
- Finucane HK, Reshef YA, Anttila V, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet*. 2018;50(4):621–9.
- Fonseca MAS, Haro M, Wright KN, et al. Single-cell transcriptomic analysis of endometriosis. *Nat Genet*. 2023;55(2):255–67.
- Tan Y, Flynn WF, Sivajothi S, et al. Single-cell analysis of endometriosis reveals a coordinated transcriptional programme driving immunotolerance

- and angiogenesis across eutopic and ectopic tissues. *Nat Cell Biol.* 2022;24(8):1306–18.
35. Liu L, Han F, Du N, et al. New insights into the ferroptosis and immune infiltration in endometriosis: a bioinformatics-based analysis. *Front Immunol.* 2024;15:1507083.
36. Guo SW. Cancer driver mutations in endometriosis: variations on the major theme of fibrogenesis. *Reprod Med Biol.* 2018;17(4):369–97.
37. Yilmaz BD, Bulun SE. Endometriosis and nuclear receptors. *Hum Reprod Update.* 2019;25(4):473–85.
38. Su Z, Su W, Li C, et al. Identification and immune characteristics study of pyroptosis-related genes in endometriosis. *Biochem Genet.* 2024;62(4):2810–29.
39. Papadatou V, Tologkos S, Tsolou A, et al. CYLD expression in endometrial carcinoma and correlation with clinicohistopathological parameters. *Taiwan J Obstet Gynecol.* 2022;61(4):596–600.
40. Correction to: AURKA inhibits the decidualization of the eutopic endometrium in endometriosis through nuclear factor-kappaB p65. *Biol Reprod.* 2025;112(6):1304–1306.
41. Kruse R, Krantz J, Barker N, et al. Characterization of the CLASP2 protein interaction network identifies SOGA1 as a microtubule-associated protein. *Mol Cell Proteomics.* 2017;16(10):1718–35.
42. Combs TP, Snell-Bergeon JK, Maahs DM, et al. Adiponectin-SOGA dissociation in type 1 diabetes. *J Clin Endocrinol Metab.* 2015;100(8):E1065–1073.
43. Matsuoka R, Miki M, Mizuno S, Ito Y, Yamada C, Suzuki A. MTCL2 promotes asymmetric microtubule organization by crosslinking microtubules on the Golgi membrane. *J Cell Sci.* 2022. <https://doi.org/10.1242/jcs.259374>.
44. Ferreira LT, Logarinho E, Macedo JC, Maia ARR, Maiato H. SOGA1 and SOGA2/MTCL1 are CLASP-interacting proteins required for faithful chromosome segregation in human cells. *Chromosome Res.* 2021;29(2):159–73.
45. Cowherd RB, Asmar MM, Alderman JM, et al. Adiponectin lowers glucose production by increasing SOGA. *Am J Pathol.* 2010;177(4):1936–45.
46. Sarsenova M, Lawarde A, Pathare ADS, et al. Endometriotic lesions exhibit distinct metabolic signature compared to paired eutopic endometrium at the single-cell level. *Commun Biol.* 2024;7(1):1026.
47. Zhang M, Xu T, Tong D, et al. Research advances in endometriosis-related signaling pathways: a review. *Biomed Pharmacother.* 2023;164:114909.
48. Michaelis UR. Mechanisms of endothelial cell migration. *Cell Mol Life Sci.* 2014;71(21):4131–48.
49. Gazvani R, Templeton A. Peritoneal environment, cytokines and angiogenesis in the pathophysiology of endometriosis. *Reproduction.* 2002;123(2):217–26.
50. Kamrani S, Amirchaghmaghi E, Ghaffari F, Shahhoseini M, Ghaedi K. Altered gene expression of VEGF, IGFs and H19 lncRNA and epigenetic profile of H19-DMR region in endometrial tissues of women with endometriosis. *Reprod Health.* 2022;19(1):100.
51. Gargett CE, Schwab KE, Deane JA. Endometrial stem/progenitor cells: the first 10 years. *Hum Reprod Update.* 2016;22(2):137–63.
52. Tissot M, Lecointre L, Faller E, Afors K, Akladios C, Audebert A. Clinical presentation of endometriosis identified at interval laparoscopic tubal sterilization: prospective series of 465 cases. *J Gynecol Obstetr Human Reprod.* 2017;46(8):647–50.
53. Khan KN, Fujishita A, Kitajima M, Hiraki K, Nakashima M, Masuzaki H. Occult microscopic endometriosis: undetectable by laparoscopy in normal peritoneum. *Hum Reprod.* 2014;29(3):462–72.
54. Gratton SM, Choudhry AJ, Vilos GA, et al. Diagnosis of endometriosis at laparoscopy: a validation study comparing surgeon visualization with histologic findings. *J Obstet Gynaecol Can.* 2022;44(2):135–41.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.