Liu *et al. Journal of Ovarian Research* (2025) 18:33
https://doi.org/10.1186/s13048-025-01617-2

Journal of Ovarian Research

RESEARCH Open Access

Polycystic ovary syndrome and epithelial-mesenchymal transition: Mendelian randomization and single-cell analysis insights

Dong Liu[1,2], Dan Liu[2,3] and Kunyan Zhou[1,2*]

Abstract

Background The process of epithelial-mesenchymal transition (EMT) may promote fibrosis in ovarian tissue related to polycystic ovary syndrome (PCOS), thus affecting ovarian function and hormonal balance.

Objective This study aimed to explore key genes associated with EMT in PCOS and their potential molecular regulatory mechanisms, exclusively from the perspective of transcriptomics and single-cell RNA sequencing (scRNA-seq), combined with Mendelian Randomization (MR) analysis.

Methods The dataset for PCOS and EMT-related genes (EMT-RGs) were sourced from public databases. The key genes in this study were identified via differential expression analysis, MR, and evaluation of expression levels. Enrichment analysis and a series of functional analyses were conducted on these genes to further elucidate their potential mechanisms. Subsequently, using scRNA-seq data and validation of the expression of key genes, key cell group in PCOS were identified, followed by pseudo-time and cell communication analyses to provide deeper insights.

Results Three key genes, NUCB2 [odds ratio (OR) = 0.8634, 95% confidence interval (CI): 0.8145–0.9152, $P < 0.0001$], PGF (OR = 0.8393, 95% CI: 0.7185–0.9805, $P < 0.05$), and CRIM1 (OR = 0.7539, 95% CI: 0.6556-0.670, $P < 0.0001$), were identified as having a unidirectional causal association with PCOS and were associated with a reduced risk of PCOS. In public datasets, NUCB2 exhibited significantly increased expression in PCOS samples, while PGF and CRIM1 showed the opposite trends. These three genes were enriched in pathways related to cellular functions, metabolic processes, and the operation of the nervous system, and they were co-expressed in smooth muscle. Additionally, five cell clusters were annotated, among which fibroblasts were identified as key cells due to their highest expression of all three key genes. Further analysis revealed a bifurcation event occurring during the mid-development stage of fibroblasts, with PCOS samples displaying a higher abundance of fibroblasts. In PCOS samples, fibroblasts exhibited more extensive communication with secretory epithelial cells, indicating a more complex intercellular interaction within this condition.

*Correspondence:
Kunyan Zhou
zhoukunyan2006@126.com

Full list of author information is available at the end of the article

Conclusion This study identified three EMT-RGs: NUCB2, PGF, and CRIM1, which were associated with a reduced risk of PCOS, with fibroblast identified as a key cell group in the disease's pathology. This provides new insights for PCOS research.

Keywords Polycystic ovary syndrome, Epithelial-mesenchymal transition, Mendelian randomization, Single-cell RNA sequencing

Introduction

Polycystic ovary syndrome (PCOS) represents the most prevalent endocrinopathy affecting adolescent females and women of reproductive age, with global prevalence rates varying from 6 to 21%, depending on the country and diagnostic criteria used [1]. Common treatments include lifestyle changes, physical activity, and medication such as combined oral contraceptives, metformin, and clomiphene [2, 3]. PCOS is characterized by hyperandrogenism, ovulatory dysfunction, and polycystic ovarian morphology, often accompanied by insulin resistance and obesity [4]. Additionally, PCOS increases the risk of several complications, including endometrial cancer, pregnancy complications (e.g., gestational diabetes and preeclampsia), cardiovascular diseases, type 2 diabetes mellitus, metabolic syndrome, depression, and anxiety [5]. Current evidence suggests a complex interaction between genetic, hormonal, and environmental factors in PCOS, though the precise etiology and pathology remain incompletely understood [6].

Epithelial-mesenchymal transition (EMT) is a crucial biological process where epithelial cells transition to a mesenchymal phenotype, gaining enhanced migratory and invasive properties [7]. This process is fundamental in embryogenesis, tissue repair, and fibrosis [8]. Recent studies have begun to explore the connection between EMT and PCOS. Evidence suggests EMT may contribute to the pathogenesis of PCOS by promoting the transformation of epithelial ovarian cells into mesenchymal cells, thereby enhancing fibrosis and disrupting normal ovarian function [7]. In PCOS, an imbalance in EMT-regulating factors such as TGF-β and inflammatory cytokines may contribute to excessive fibrotic tissue formation in the ovaries, leading to impaired ovarian function and folliculogenesis [7, 9]. A study has revealed that ALG2 plays a critical role in ovarian functions by showing downregulation under hypoxic conditions and inhibiting EMT and stemness of ovarian granulosa cells through suppressing the Wnt/β-catenin signaling pathway [10]. Disrupted regulation of EMT and MAPK signaling pathways may result in impaired endometrial cell homeostasis and functionality, thereby reducing the reproductive potential of PCOS patients [11, 12]. Research indicates that hyperandrogenism in PCOS can exacerbate EMT processes, further contributing to ovarian dysfunction and the characteristic polycystic ovarian morphology [13]. Despite numerous studies investigating various molecular mechanisms in PCOS patients, the expression of EMT-related regulators in PCOS have not been extensively explored.

Mendelian randomization (MR) is a method that utilizes genetic variants as instrumental variables to assess causal relationships between risk factors and diseases [14]. This approach typically leverages data from genome-wide association studies (GWAS). MR helps determine causality by minimizing confounding and reverse causation biases, making it a robust tool compared to traditional observational studies [15]. Numerous MR studies have shown that immune cells, gut microbiota [16], age at menarche (>15 years), age at menopause, obesity, testosterone levels, fasting insulin levels, and depression appear to play causal roles in the etiology of PCOS [17–19]. Additionally, studies have found that PCOS can act as a risk factor for other diseases and is causally associated with an increased risk of breast cancer and chronic kidney disease [19, 20]. MR effectively controls for confounding factors, making it crucial for identifying the causal relationship between differentially expressed EMT-related genes (DE-EMT-RGs) and PCOS. This method enhances the reliability of results and offers valuable insights into the underlying mechanisms of PCOS [19].

Single-cell RNA sequencing (scRNA-seq) is an advanced technology that enables high-throughput, multidimensional analysis at the individual cell level. Unlike traditional RNA sequencing, which averages transcript levels across a cell population, scRNA-seq provides complete transcriptional profiles of individual cells [21]. This allows for detailed analysis of intracellular and intercellular interactions, identification of cellular heterogeneity, discovery of new cell types, and detection of dynamic changes in cell states. The improved resolution of scRNA-seq facilitates the identification of potential therapeutic targets and characterizes cellular and molecular profiles of disease [22]. Study has utilized scRNA-seq technology to analyze the mechanisms of PCOS-related hyperandrogenism and predict potential therapeutic targets [23]. Additionally, single-cell sequencing combined with transcriptome analysis has revealed that abnormal mitochondrial function in oocytes at the germinal vesicle stage may contribute to decreased oocyte quality in patients with PCOS [24]. Combining single-cell sequencing with Mendelian randomization enables high-resolution analysis of gene expression at the single-cell level, reduces confounding factors and reverse causality bias,

and more accurately identifies causal mechanisms of diseases and predicts potential therapeutic targets.

In this study, we integrated transcriptome data of PCOS from public databases with single-cell sequencing data and employed Mendelian randomization analysis to identify key EMT-related genes that are causally associated with PCOS. This approach provided novel insights into the relationship between PCOS and EMT, as well as the underlying regulatory mechanisms.

Methods

Data source

To obtain transcriptome datasets related to Polycystic Ovary Syndrome (PCOS), two separate datasets were retrieved from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/gds): GSE155489 and GSE168404. Dataset GSE155489, based on the GPL20795 platform, included samples from 5 PCOS patients and 5 normal control granulosa cells, and was designated as the training set [25]. On the other hand, dataset GSE168404, based on the GPL16791 platform, comprised samples from 5 PCOS patients and 5 normal control granulosa cells, serving as the validation set [26].

Additionally, a single-cell RNA sequencing (scRNA-seq) dataset was selected from an early published article and was available at https://zenodo.org/record/7942968 [23], comprising 20 ovarian tissue samples from 5 PCOS patients and 5 controls. Notably, the dataset was derived from whole ovarian tissues, containing a heterogeneous cell population, including theca cells, fibroblasts, immune cells, and secretory epithelial cells, reflecting the complex ovarian microenvironment.

Moreover, 1,185 EMT-RGs were obtained from the dbEMT 2 database (http://dbemt.bioinfo-minzhao.org/download.cgi), providing a comprehensive resource for exploring the role of EMT in PCOS pathophysiology.

Selection and enrichment analysis of differentially expressed genes (DEGs) and differentially expressed EMT-RGs (DE-EMT-RGs)

Differential expression analysis was conducted between PCOS and control samples in the training dataset using the DESeq2 package (v 1.34.0) [27] to screen for DEGs ($|\log_2 FC| \geq 0.5$ and $P < 0.05$). Visualization was performed using the ggplot2 (v 3.3.5) [28] and pheatmap packages (v 1.0.12) [29] to create volcano plot and heatmap, respectively. Subsequently, these DEGs were intersected with 1,185 EMT-RGs from an existing database via ggvenn package (v 1.7.3) [30] to identify DE-EMT-RGs specific to this study. Further analysis was conducted on these DEGs and DE-EMT-RGs, utilizing the clusterProfiler package (v 4.6.0) [31] for Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses. The GO analysis encompassed three categories: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF). Visualization of the enrichment analysis was achieved using ggplot2. Following that, in order to further explore the relationships of DE-EMT-RGs at the protein level, these DE-EMT-RGs were input into a search tool for the retrieval of interacting genes (STRING) database (http://string-db.org) with a set interaction score of greater than 0.7. The protein-protein interaction (PPI) network was then visualized using Cytoscape software (v 3.9.1) [32].

Identification of causal exposure factors associated with PCOS through Mendelian randomization (MR) analysis

To investigate the causal link between DE-EMT-RGs and PCOS, we employed MR analysis leveraging data from two distinct datasets: the 'DE-EMT-RGs' dataset and the 'PCOS' dataset (finngen_R9_E4_PCOS), as exposure factors and outcome variable respectively. The Expression Quantitative Trait Loci (eQTL) data for 'DE-EMT-RGs' as well as the 'PCOS' dataset (finngen_R9_E4_PCOS) utilized genome-wide association study (GWAS) identifiers from the FinnGen database (https://www.finngen.fi/en/access_results). A successful MR analysis hinged on three critical prerequisites: (1) the presence of a robust and significant association between the instrumental variables (IVs) and the exposure, (2) the independence of the IVs from any confounding variables, and (3) the specificity of IVs affecting the outcome solely through the exposure.

Utilizing the TwoSampleMR package (v 0.5.8) [33], our MR study commenced with the extract_instruments function, which identified single nucleotide polymorphisms (SNPs) closely linked to the exposure factor, marked by a significance threshold ($P < 5 \times 10^{-6}$). We excluded any SNPs in linkage disequilibrium (LD) (using a 10,000 kb window and a threshold of $R^2 < 0.001$) and those strongly associated with the outcome. Afterward, the mv_harmonise_data function was enlisted to ensure coherence in effect alleles and sizes across the datasets, selectively filtering and refining the SNPs to fulfill the stringent criteria essential for a valid MR analysis. This thorough approach reinforced the robustness of our findings, aiming to conclusively illuminate any causal connections between DE-EMT-RGs and PCOS. MR analysis was conducted by employing mr function in combination with five algorithms: MR-Egger [34], Weighted median [35], Inverse variance weighted (IVW) [36], Simple mode [33] and Weighted mode [37]. In this study, the IVW method served as the primary reference for interpreting the results of the MR analysis ($P < 0.05$). Odds Ratio (OR) > 1 suggested that the exposure factor was a risk factor for the outcome variable, whereas OR < 1 indicated that the exposure factor was a protective factor for the outcome variable.

Recognition of key genes with causal association to PCOS

Thereafter, we evaluated the expression levels of exposure factors with causal association to PCOS in both the training and validation datasets. Genes that exhibited significant differences and consistent expression trends across these datasets were defined as key genes in this study ($P < 0.05$). For these genes, scatter plots, forest plots, and funnel plots were created to visualize the MR results. To ascertain the robustness of the MR findings, a comprehensive suite of sensitivity analyses was conducted. The assessment began with an evaluation of heterogeneity using Cochran's Q test via the mr_heterogeneity function; a *P*-value greater than 0.05 indicated an absence of significant heterogeneity [38]. Subsequent analysis included testing for horizontal pleiotropy, which might indicate potential confounders. This was performed using both the mr_pleiotropy_test and Mrpresso functions, where a *P*-value greater than 0.05 suggested no noticeable pleiotropy [39]. Additionally, the Leave-One-Out (LOO) sensitivity analysis was implemented, whereby individual SNPs were iteratively removed to verify the stability of the results when excluding specific variants [40]. To further ensure the reliability of the causal inference, the Steiger directional test was applied to address the possibility of reverse causation. Confirmation of the causal association was strengthened when the correct_causal_direction outcome was identified as TRUE and the Steiger *P*-value was lower than 0.05 [41].

Functional and annotation analyses

In an effort to further explore the signaling pathways enriched with key genes, in the training dataset, the psych package (v 2.2.9) [42] was utilized to perform Spearman correlation analysis between each key gene and all other genes in the samples. The genes were then sorted based on $\log_2$FC. Using the clusterProfiler package and KEGG gene sets from the Molecular Signatures Database (MSigDB) (https://www.gsea-msigdb.org/gsea/msigdb), Gene Set Enrichment Analysis (GSEA) was conducted [|Normalized Enrichment Score NES)|>1 and adj. $P < 0.05$]. Moreover, Gene Set Variation Analysis (GSVA) was performed to scrutinize differences in enriched pathways between PCOS and control samples. Initially, the pathway scores for PCOS samples were calculated in the training set using the GSVA package (v 1.42.0) [43]. Next, the PCOS samples were stratified into high and low expression groups according to the median expression levels of key genes. Differences in the GSVA scores between these groups were then analyzed using the limma package (v 3.50.1) [44] ($P < 0.05$).

Comprehensive analysis of key genes: subcellular localization, chromosomal positions, functional associations, and differential expression across immune cells and tissues

Differential expression of key genes across various immune cells could significantly impact their functionality. Consequently, using the Human Protein Atlas (HPA) database (https://www.proteinatlas.org/), the expression of these key genes in 18 different immune cell types was analyzed. This comprehensive analysis aimed to elucidate the distinct roles these genes played in immune cell regulation and their broader implications in immunology. After that, to investigate the subcellular localizations of key genes, we employed the Hum-mPLoc 3.0 (http://www.csbio.sjtu.edu.cn/bioinf/Hum-mPLoc3/), a tool designed to analyze the subcellular localization of proteins encoded by these genes. In addition, to elucidate the specific chromosomal locations of these key genes, the RCircos package (v 1.2.2) [45] was utilized, and this software aided in visualizing the chromosomal positions of genes, providing valuable insights into their genomic arrangement. Then, to explore potential genes associated with the function of our key genes, we turned to the GeneMANIA database (http://www.genemania.org). This powerful resource offered predictions on genes that may be functionally similar or related, enriching our understanding of gene networks. Eventually, to ascertain the mRNA expression patterns of the key genes across various organs and tissues, we used the BioGPS database (http://biogps.org). This platform facilitated the vis ualization and analysis of gene expression data, enabling us to explore the differential expression of these genes in diverse biological contexts.

Construction of regulatory network

Subsequent analyses aimed to elucidate the molecular regulatory mechanisms of key genes by constructing regulatory networks. Initially, potential miRNAs regulating key genes were predicted using the Starbase database (https://rnasysu.com/encori/), selecting those with a clipExpNum greater than 20. Further predictions were made using the miRDB database (https://mirdb.org/), applying a threshold of Target Score greater than 60. The intersection of miRNAs from both databases yielded the final set of miRNAs. The corresponding lncRNAs for these final miRNAs were then determined using the StarBase database, followed by the construction of a lncRNA-miRNA-mRNA network. Concurrently, transcription factors (TFs) targeting the key genes were obtained from the ChIP-X Enrichment Analysis 3 (ChEA3) database (https://maayanlab.cloud/chea3/), with the selection criterion being a Rank less than 200. This information facilitated the creation of a TF-mRNA network. Visualization of the results was accomplished using Cytoscape

software. In addition, the SIGNOR database (https://signor.uniroma2.it/) was employed for annotating proteins of the key genes, allowing for the construction of a key gene signal information network that focused on exploring the signal transduction relationships of the key genes.

Potential drug targets and molecular docking studies for the treatment of PCOS

A series of analyses were conducted on potential drugs targeting key genes for the treatment of PCOS. In the beginning, the Drug-Gene Interaction database (DGIdb) (http://dgidb.genome.wustl.edu/) was used to predict potential drugs targeting key genes. Afterward, to further analyze these predictions, we employed AlphaFold v2.0 (https://alphafold.ebi.ac.uk/) as an artificial intelligence tool to predict the 3D structures of the key genes. The amino acid sequences of the key genes were downloaded from the UniProt-KB database (http://www.uniprot.org/). The structural integrity of the 3D models was assessed using the Local Distance Difference Test (LDDT) scores from the AlphaFold database, with scores above 90 indicating excellent stereochemical performance of the key gene models.

Next, molecular docking studies were established, specifically the 3D structures of target drugs were retrieved and downloaded in SDF format from the PubChem database (https://pubchem.ncbi.nlm.nih.gov/). Likewise, the 3D structures of the key gene proteins were retrieved and downloaded in PDB format from the UniProt database (https://www.uniprot.org/). Molecular docking was then executed on the CB-DOCK2 database (http://cadd.labshare.cn/cb-dock2/index.php). The results for five different active sites (CurPocket_ID) were presented, including Vina docking scores, cavity volumes, spatial center coordinates, and the sizes of the docking regions. The Vina scores indicated the binding affinity of small molecules at these sites, with binding energies less than −5 kJ/mol suggesting strong binding capabilities.

ScRNA-seq analysis

The comprehensive analysis of the scRNA-seq dataset was analyzed by Seurat package (v 4.1.0) [46]. Originally, a meticulous data quality control (QC) process was implemented on the acquired samples from 5 patients with PCOS and 5 control individuals within the scRNA-seq dataset. The objective was to excise any low-quality data attributable to cellular deterioration or unsuccessful library preparations. The established exclusionary criteria encompassed the elimination of genes with detectable expression in less than 200 cells, the removal of cells exhibiting a nFeature_RNA (the count of distinct genes) in excess of 8000, a nCount_RNA (total gene expression counts) surpassing 80,000, as well as discarding cells with a mitochondrial gene expression percentage (percent.mt) of 10% or higher. Subsequently, leveraging the JackStrawPlot and JackStraw functions, we diligently pinpointed statistically robust principal components (PCs). These components served as a foundation for dimensionality reduction through principal component analysis (PCA). Following this reduction, we embarked on an unsupervised clustering analysis of the cells employing the FindNeighbors and FindClusters functions, setting the resolution parameter to 1. The clustering results were then depicted using the uniform manifold approximation and projection (UMAP) for visual representation. In the final phase of our analysis, cell clusters were annotated by referencing the distinct marker genes catalogued in the CellMarker2 database (http://bio-bigdata.hrbmu.edu.cn/CellMarker/), and the distribution of marker genes in each cell cluster was displayed.

Identifying key cells and constructing cell communication and pseudo-time analyses

After annotating multiple cell clusters, we constructed a series of analyses to further investigate at the cellular level. At the outset, we explored key cells in PCOS by assessing the expression of key genes within these clusters. Cell clusters exhibiting the highest expression levels of these key genes were defined as key cells for this study. Subsequently, to further explore the differentiation trajectories and evolutionary paths of these key cells during development, the reduceDimension function in the Monocle3 package (v 2.26.0) [47] combined with the DDRTree algorithm was utilized to perform dimensionality reduction. This step simplified the complex relationships between cells, making it more suitable for visualization and further analysis. Next, we used the order Cells function to sort the key cells based on the reduced dimensionality data, inferring their pseudotime trajectories in biological processes. Ultimately, for the identified cell clusters, we conducted communication analysis using the CellChat package (v 1.5.0) [48]. This involved calculating the communication likelihood at the signaling pathway level by computing all ligand-receptor interactions associated with each pathway and constructing ligand-receptor networks. To better infer the node sizes and edge weights in the networks across different groups, we calculated the maximum number of interactions and interaction weights for each cell in both control and PCOS samples.

Statistical analysis

All analyses were executed in R software (v 4.2.2). Differences between groups were analyzed by the Wilcoxon test. $P < 0.05$ was considered statistically significant.

Results

Identification of 215 DE-EMT-RGs enriched in multiple functions and pathways

Based on differential expression analysis between PCOS and control samples in the training dataset, 3,114 DEGs were gained, including 1,559 upregulated genes and 1,555 downregulated genes (Fig. 1a). Subsequent enrichment analysis of the DEGs characterized 240 GO terms, comprising 166 BP, 66 CC, and 8 MF. These terms were related to metabolic processes, cell adhesion, and signaling, as well as the structure and function of organelles. In the KEGG pathway analysis, 80 pathways were enriched, covering amino acid metabolism, energy production, disease mechanisms, and signaling pathways (Fig. 1b). Intersection between these 3,114 DEGs and 1,185 EMT-RGs resulted in 215 DE-EMT-RGs (Fig. 1c). Enrichment

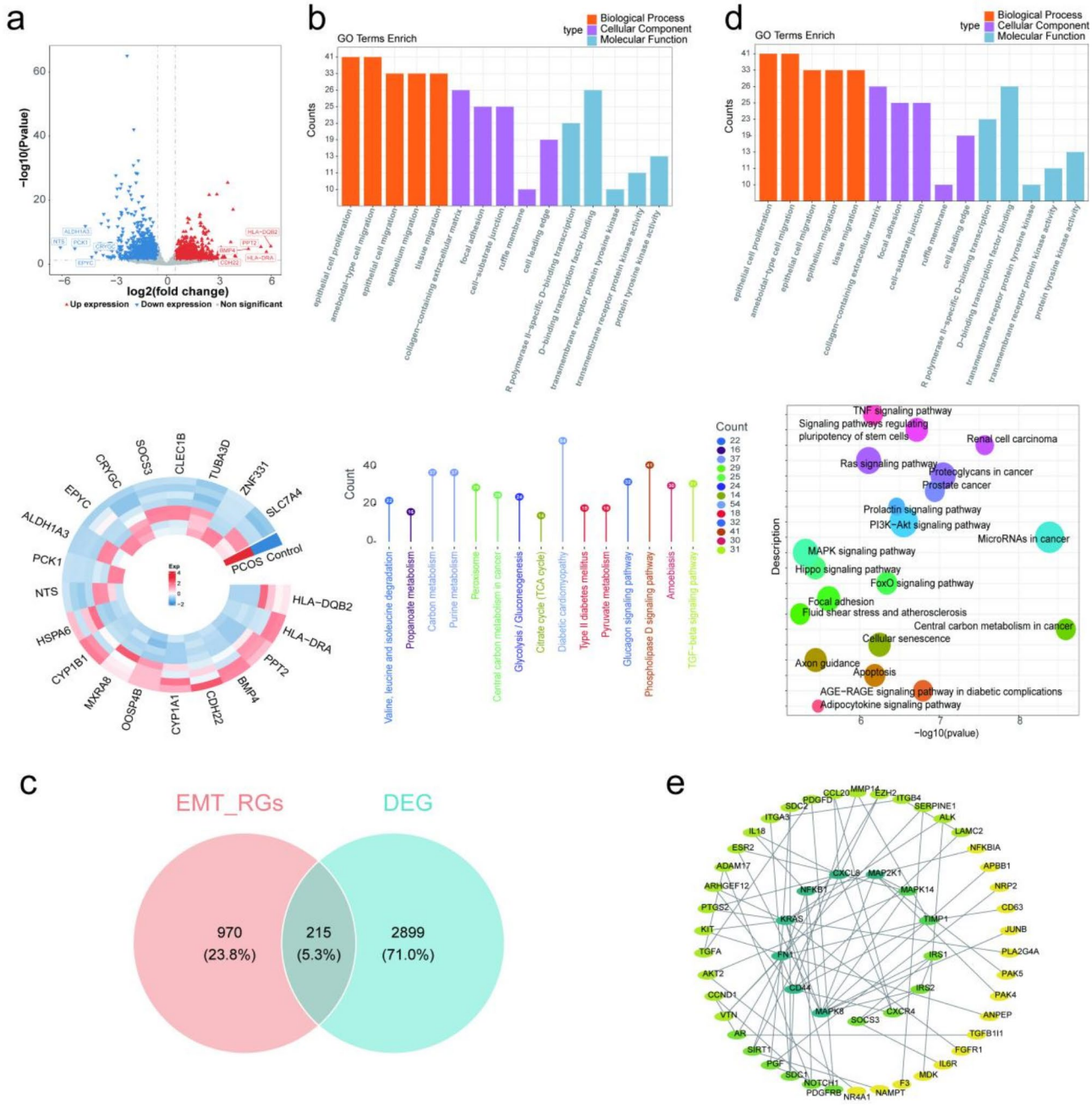

Fig. 1 Identification and enrichment analysis of differential genes. (**a**-1,2) Volcano plot of differential gene analysis. Red dots represent upregulated genes, blue dots represent downregulated genes, and grey dots represent genes with no significant difference or smallfold changes. (**b**-1,2) GO and KEGG Enrichment Analysis of Differentially Expressed Genes (DEGs). (**c**) Venn map to obtain a total of 215 candidate genes shared by EMT-RGs and DEGs. (**d**-1,2) Enrichment pathway network using the key KEGG results. (**e**) PPI network constructed using the 215 DE-EMT-RGs

analysis of these genes highlighted 1,749 GO entries, including 1,665 BPs, 42 CCs, and 42 MFs, mainly dealing with cell migration and adhesion, signaling pathways, and mechanisms of gene expression regulation. These were closely associated with physiological and pathological processes such as cancer metastasis, wound healing, and tissue remodeling. The KEGG analysis identified 124 pathways that span across aspects like cancer biology, metabolic diseases, cell signaling, cell fate determination, and cell stress responses, revealing crucial insights on molecular regulation of cell and tissue behavior, pivotal for understanding human health, disease, and therapeutic interventions (Fig. 1d). Furthermore, a PPI network constructed using the 215 DE-EMT-RGs included 95 nodes with 40 edges, an average node degree of 0.884, and an average local clustering coefficient of 0.271 (Fig. 1e). Notably, proteins such as KRAS, FN1, CD44, MAPK8, NFKB1, and CXCL8 showed high interaction scores, indicating significant roles in the network's functionality.

Key genes NUCB2, PGF, and CRIM1 as protective factors for PCOS

Subsequently, we took 215 DE-EMT-RGs as exposure factors and PCOS as the outcome variable, aiming to identify genes that had a causal association with PCOS. After selecting IVs, the IVW method revealed 55 exposure factors that were significantly causally associated with PCOS ($P<0.05$) (Supplementary Table 1). Further expression validation analysis was conducted to refine the screening. In both the training and validation sets, the expressions of these 55 genes were examined (Fig. S1). Unfortunately, expression data for CXCL8 was not found in the validation set, hence expressions of only 54 genes were presented in that set. The results showed that only NUCB2, PGF, and CRIM1 had consistent and significantly different expression trends in the two training sets ($P<0.05$), with NUCB2 exhibiting a notably lower expression trend in PCOS samples ($P<0.05$), while PGF and CRIM1 showed the opposite ($P<0.05$). Interestingly, in the MR analysis, NUCB2 (OR = 0.8634, 95% confidence interval (CI): 0.8145–0.9152, $P<0.0001$), PGF (OR = 0.8393, 95% CI: 0.7185–0.9805, $P<0.05$), and CRIM1 (OR = 0.7539, 95% CI: 0.6556-0.670, $P<0.0001$) all emerged as protective factors for PCOS. The underlying mechanisms of these relationships warranted further discussion.

The scatter plots depicted a striking negative correlation between three genes and PCOS, underlining their role as protective factors against PCOS, devoid of confounding variables (Fig. 2a). Meanwhile, the forest plots demonstrated that these genes exerted a positive influence on PCOS protection, with fixed effects showing a value less than zero, implying that these genes might also decrease the probability of developing PCOS (Fig. 2b). The methodological integrity of the study was corroborated by the funnel plot, which confirmed compliance with the second axiom of MR (Fig. 2c).

Additionally, we conducted a sensitivity analysis to further confirm the reliability of the MR analysis. Specifically, the absence of heterogeneity in the sample was indicated by the results of Cochran's Q test ($P>0.05$) (Supplementary Table 2), and the horizontal pleiotropy test and Mrpresso further demonstrated that there was no horizontal pleiotropy in the MR study (mr_heterogeneity and Mrpresso $P>0.05$) (Supplementary Table 3). Following the systematic removal of individual SNPs, the impact attributed to the residual SNPs on the outcome variables remained notably stable, thereby reaffirming the robustness of the Mendelian Randomization analysis outcomes (Fig. 2d). At last, the Steiger test was conducted, further substantiating the unidirectional causal association between NUCB2, PGF, and CRIM1 and PCOS (correct_causal_direction = TRUE and steiger_p val < 0.05) (Supplementary Table 4).

Deciphering the role of key genes and pathways in PCOS pathogenesis

Further analysis, utilizing the combination of GSEA and GSVA, aimed to explore some signaling pathways in key genes as well as between PCOS and control samples. Particularly, three key genes were commonly enriched in the "Parkinson's disease" and "proteasome" pathways. Additionally, PGF and CRIM1 were also co-enriched in pathways such as "glycolysis gluconeogenesis", "ribosome", and "axon guidance" (Fig. 3a-c). These pathways were intricately linked to cellular function, metabolic processes, and neural system operations, suggesting that these key genes could potentially impact the onset and progression of PCOS by modulating these critical biological processes.

GSVA results highlighted notable differences between PCOS and control samples in several GO terms, including "reactome ionotropic activity of kainate receptors", "reactome presynaptic depolarization and calcium channel opening", "reactome metabolism of polyamines", and "reactome ABC transporter disorders" (Fig. 3d). In the KEGG analysis, pathways such as "glycosphingolipid biosynthesis ganglio series", "ribosome", "circadian rhythm in mammals", and "protein export" showed significant variation, underscoring potential areas for further investigation in the pathophysiology of PCOS (Fig. 3e).

Expression, localization, and interaction of NUCB2, CRIM1, and PGF in immune response and organ-specific gene networks

A suite of analyses was carried out to deepen our understanding of key genes. The initial stage involved assessing the expression levels of these genes across 18

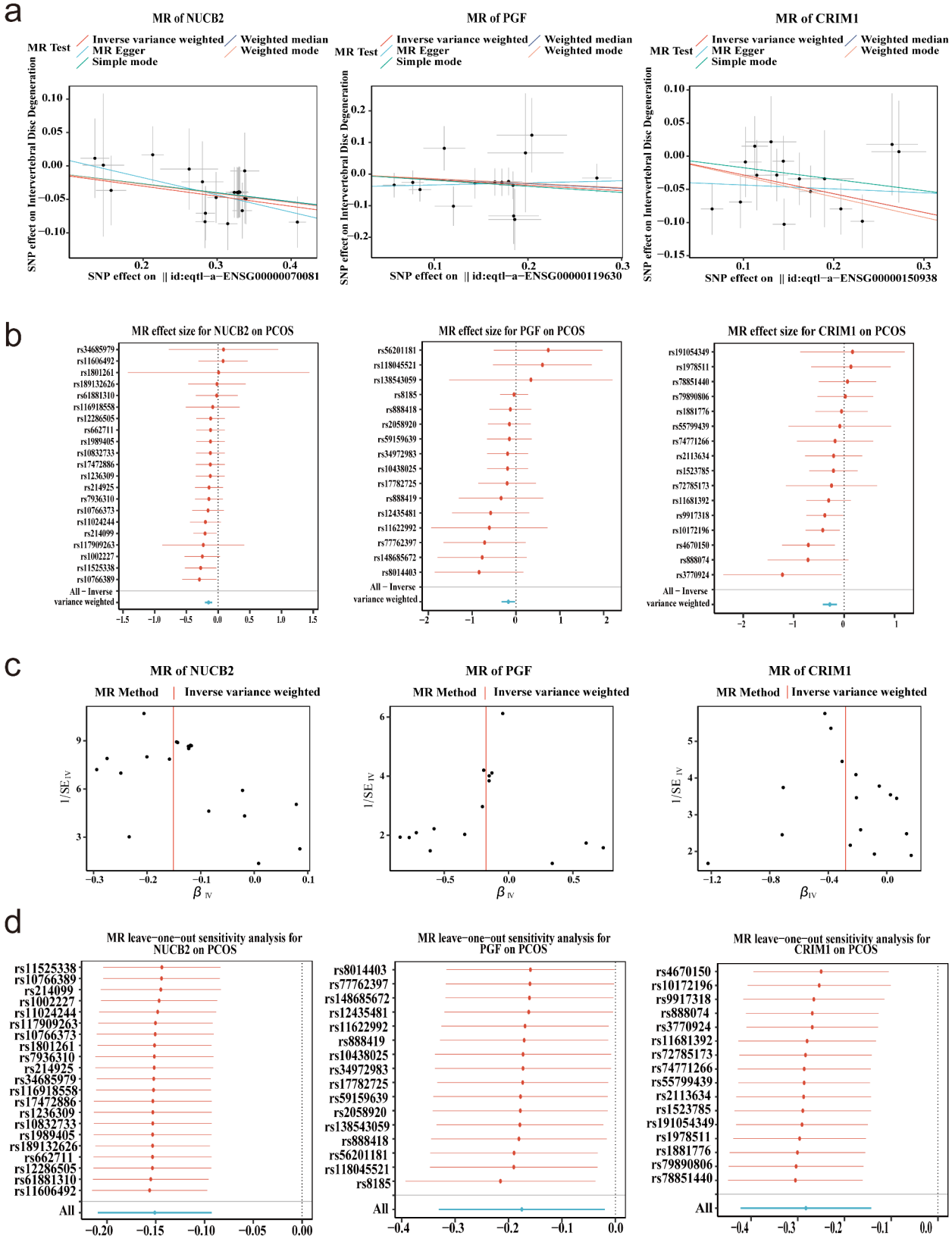

Fig. 2 Screening and identification of key genes. (**a**) Scatter Plot Analysis of the Expression Patterns of NUCB2, PGF, and CRIM1 Genes. (**b**) Forest Plot Analysis of the Association between PCOS and the Genes NUCB2, PGF, and CRIM1. (**c**) Analysis of Funnel Plots for the Genes NUCB2, PGF, and CRIM1 Genes. (**d**) Leave-One-Out Sensitivity Analysis of the Association between PCOS and the Genes NUCB2, PGF, and CRIM1

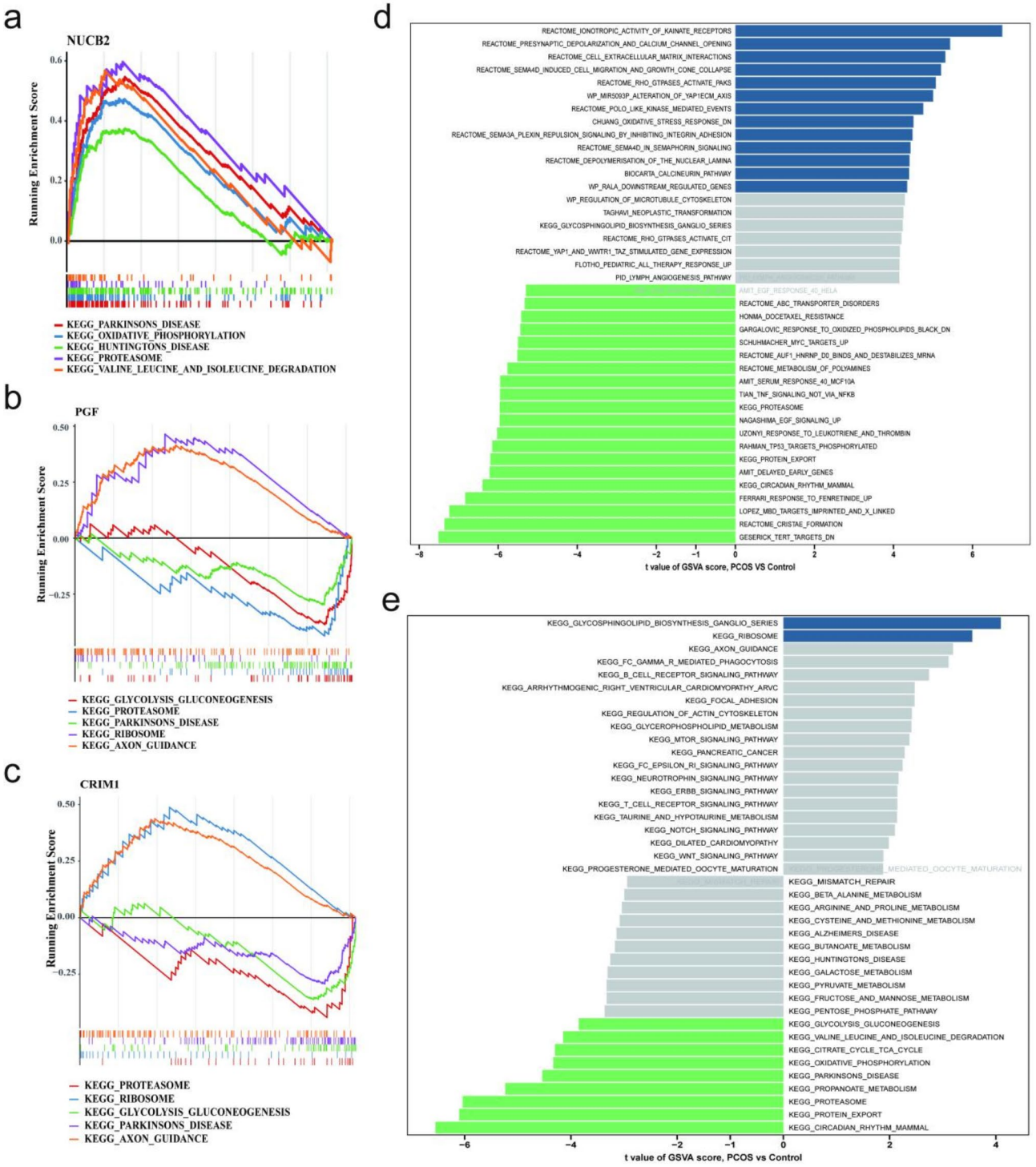

Fig. 3 Acquisition of key pathways. (**a**) KEGG Pathway Enrichment Analysis of the NUCB2 Gene. (**b**) KEGG Pathway Enrichment Analysis of the PGF Gene. (**c**) KEGG Pathway Enrichment Analysis of the CRIM1 Gene. (**d**) Key Pathway GSVA Scores in PCOS and Control Samples. (**e**) KEGG Pathway Enrichment Analysis of the Pathophysiology of PCOS

immune cell types, revealing a notably higher expression of NUCB2 in these cells, peaking in Plasmacytoid Dendritic cells (Fig. 4a). In a similar patterning, CRIM1 also demonstrated maximal expression in Plasmacytoid Dendritic cells, whereas PGF didn't exhibit any prominent expression traits (Fig. 4b, c). Further investigation into the subcellular localization of these genes uncovered their dominant cellular residence. Remarkably, CRIM1's

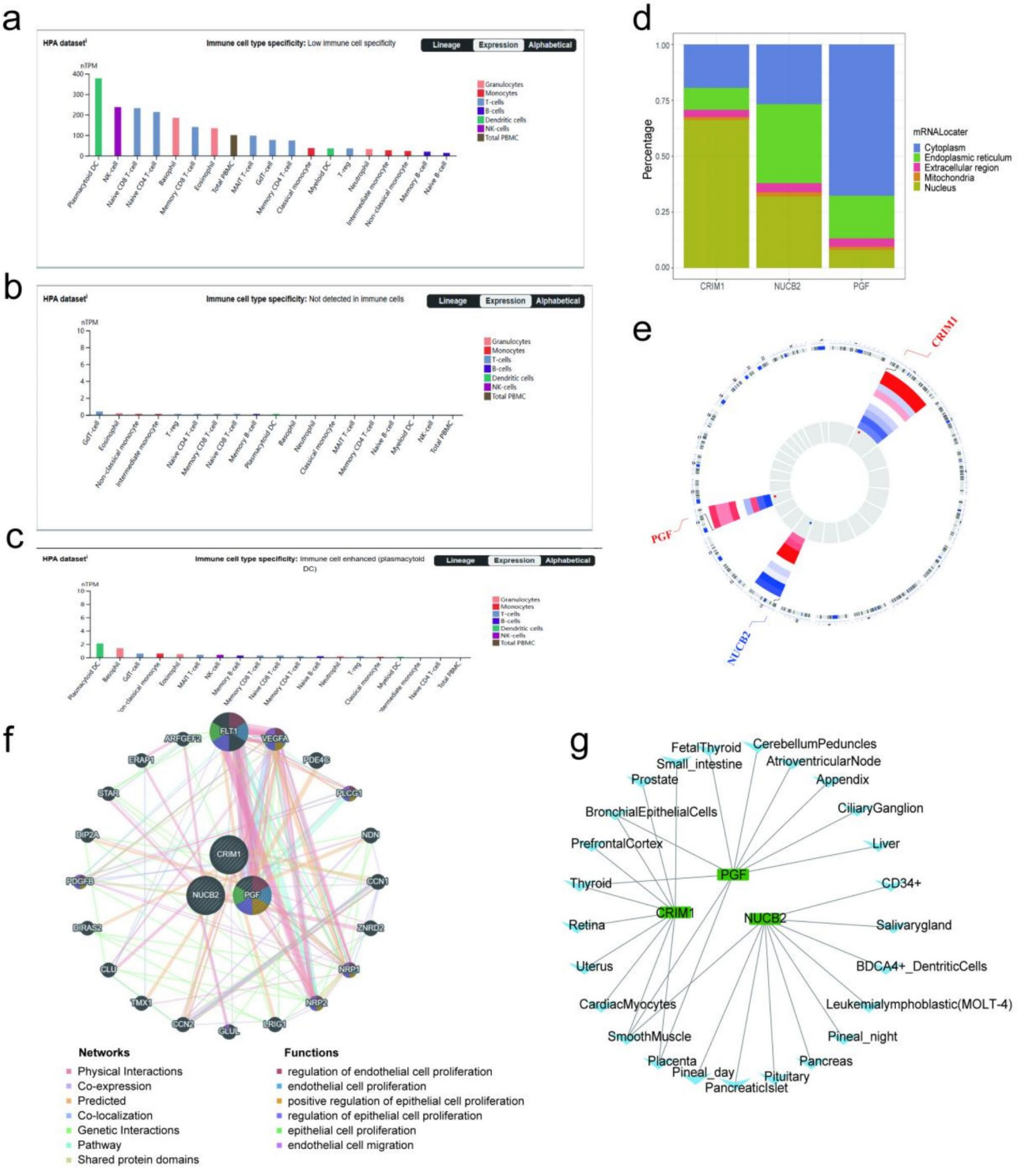

Fig. 4 Expression and localization of key Genes in immune response and organ-specific gene networks. (**a**) Expression Levels of Key Genes in Immuno-cells with Low Specificity. (**b**) Expression Levels of Key Genes in Non-Specific Immune Cells. (**c**) Expression Levels of Key Genes in Highly Specific Immune Cells.(**d**) Subcellular localization of key genes. (**e**) Chromosomal localization of key genes. (**f**) The Gene-Gene Interaction (GGI) network. (**g**) mRNA expression profiles of NUCB2, PGF, and CRIM1 genes in diverse organs and tissues

mRNA primarily occupied the nucleus, NUCB2's mRNA was predominantly distributed across the cytoplasm, endoplasmic reticulum, and nucleus, and PGF's mRNA was mainly found in the cytoplasm (Fig. 4d). Chromosomal localization studies pinpointed the positions of these genes, with CRIM1 anchored on chromosome 2, NUCB2 on chromosome 11, and PGF on chromosome 14 (Fig. 4e). Additionally, characterization of the Gene-Gene Interaction (GGI) network highlighted 20 genes tied to the functions of the trio (Fig. 4f). Nevertheless, specific functions related to NUCB2 and CRIM1 remained elusive, whereas PGF was implicated in several functions, including the regulation of endothelial cell proliferation, positive regulation of epithelial cell proliferation, and other related proliferative processes. Acknowledging the intrinsic link between a gene's function and its locational context, we leveraged the BioGPS database to map the mRNA expression profiles of these key genes across various organs and tissues. The resulting organ/tissue-gene network vividly illustrated the expression of these genes in the top 10 organs or tissues, laying out a network of 30 edges connecting the genes to 25 different organs or tissues (Fig. 4g). Noticeably, all three genes were co-expressed in smooth muscle, with PGF and CRIM1 also sharing expression sites in bronchial epithelial cells, the thyroid, and the placenta, offering a refined glimpse into the spatial dynamics of gene activity.

Elucidating the complex regulatory networks of key genes

Further investigation focused on elucidating the molecular regulatory mechanisms of key genes through the construction of regulatory networks. Within the Starbase database, a selection of 41 miRNAs along with 284 interaction pairs were pinpointed as potential regulators of key genes. The miRDB database yielded an additional 214 miRNAs and 217 interaction pairs. By intersecting these findings, we identified a core set of 51 miRNAs and their corresponding 51 interaction pairs. Predictions from the StarBase database included 41 miRNAs associated with 17 lncRNAs, comprising a total of 157 interaction pairs. This culminated in the assembly of an intricate lncRNA-miRNA-mRNA network for visualization, featuring 2 mRNAs (PGF and CRIM1), 41 miRNAs, and 17 lncRNAs, orchestrating a network with 157 complex interaction dynamics (Fig. 5a). For instance, it was suggested that XIST could modulate CRIM1 via hsa-miR-16-5p, while MALAT1 might influence PGF through hsa-miR-449a. Moreover, through the ChEA3 database, three TFs (MTF1, STAT3, and GCM1) were identified as potential regulators of PGF (Fig. 5b), further expanding our understanding of its regulatory landscape. The analysis of signal transduction relationships among key genes reveals that only the PGF protein was enriched within the signaling network (Fig. 5c). This finding suggested a potential link to Aflibercept's mechanism of action.

Exploring potential drug targets in key genes

Following drug prediction analysis, it was discovered that Aflibercept might serve as an inhibitor or binding agent for PGF, with Conbercept also potentially acting as an inhibitor of PGF (Fig. 6a). However, no evidence has been found to suggest a direct interaction between Hydrochlorothiazide and NUCB2, or that TB-403 directly interacts with PGF. Additionally, the further study presented protein models for three key genes, enabling a detailed visual inspection of their structural conformations. The 3D model for NUCB2 showcased an exceptionally high level of accuracy, with the majority of regions achieving a pLDDT score greater than 90, confirming the predicted structure's considerable reliability (Fig. 6b). In the case of PGF, while some regions achieved similarly high reliability with pLDDT scores exceeding 90, other regions scored less than 70 in pLDDT, suggesting that although some sections of the 3D model were dependable, others might be less predictable due to inherent structural flexibility or intricacy (Fig. 6c). For CRIM1, the bulk of the structure scored within a high confidence interval of $90 > \text{pLDDT} > 70$, which indicated a generally high level of precision and that the majority of the structural predictions for this region were trustworthy (Fig. 6d). In our molecular docking analysis, we ultimately obtained the 3D structures corresponding to both NUCB2 and HYDROCHLOROTHIAZIDE. The binding energy between NUCB2 and HYDROCHLOROTHIAZIDE was −6.7 kJ/mol. Within the complex, there were numerous instances, such as R188, K148, and D286, forming hydrogen bonds with HYDROCHLOROTHIAZIDE (Fig. 6e).

Identification of fibroblasts as a key cell group

After conducting QC on scRNA-seq data from 10 samples, we successfully identified a total of 22,018 cells (Fig. S2). Subsequent to the essential data processing steps, we pinpointed 2,000 highly variable genes for in-depth analysis (Fig. S3). Through PCA, which revealed no outliers or anomalous cells, we selected the PCs top 50 for further investigative procedures (Fig. S4). This led to the identification of 13 distinct cell clusters (Fig. 7a, b), with each cluster's marker gene expression levels illustrated (Fig. 7c). We successfully annotated five of these cell clusters, specifically identifying them as theca cells, dendritic cells, secretory epithelial cells, immune cells, and fibroblasts (Fig. 7d). The utilization of bubble plot underscored the marker genes' remarkable specificity, allowing us to confidently name the cells based on these marker genes (Fig. 7e). Following this, the expression of key genes within these cell clusters was further evaluated (Fig. 7f, g). It was observed that NUCB2 and CRIM1

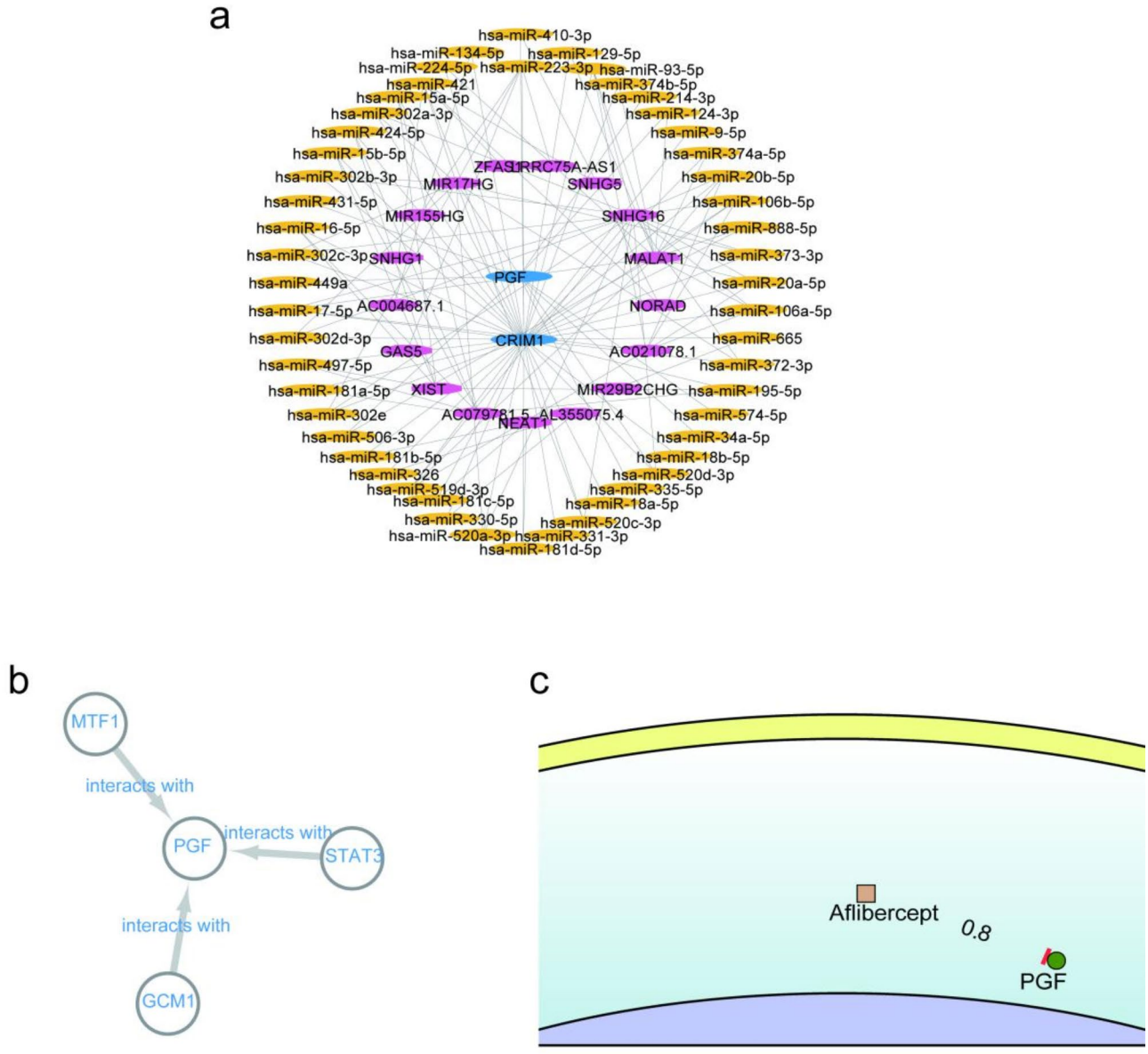

Fig. 5 Regulatory networks of key genes. (**a**) lncRNA-miRNA-mRNA interaction network. (**b**) MTF1, STAT3, and GCM1 as potential regulators of PGF. (**c**) Signal transduction among key genes

exhibited higher levels of expression within the clusters. Fibroblasts, due to their highest expression of all three key genes, were defined as the key cells in this study.

Trajectories of fibroblast maturation and heterogeneity in PCOS

Following the identification of key cells, fibroblasts, pseudo-time trajectory inference was conducted. We determined the starting point of the fibroblast trajectory, indicating that as cells moved away from this starting point, they underwent maturation during their developmental process (Fig. 8a). It was clearly observable that at both early and middle stages of cell development, the number of fibroblasts significantly surpassed their quantity in the later stages of development. Notably, a bifurcation occurred in the mid-development stage, typically signifying that at this developmental point, a single cell population began diverging into different cellular states or fates, indicating the emergence of heterogeneity within the original cell population. This heterogeneity was sufficient to guide cells along divergent developmental trajectories. Moreover, we further divided fibroblasts into six clusters, with each cluster representing a collection of cells that share similar phenotypic or functional characteristics (Fig. 8b). It could be seen that clusters 0 and 4 predominantly distributed during the early and late stages of cell development, while clusters 1 and 6 mainly distributed during the mid-stage, with clusters 5

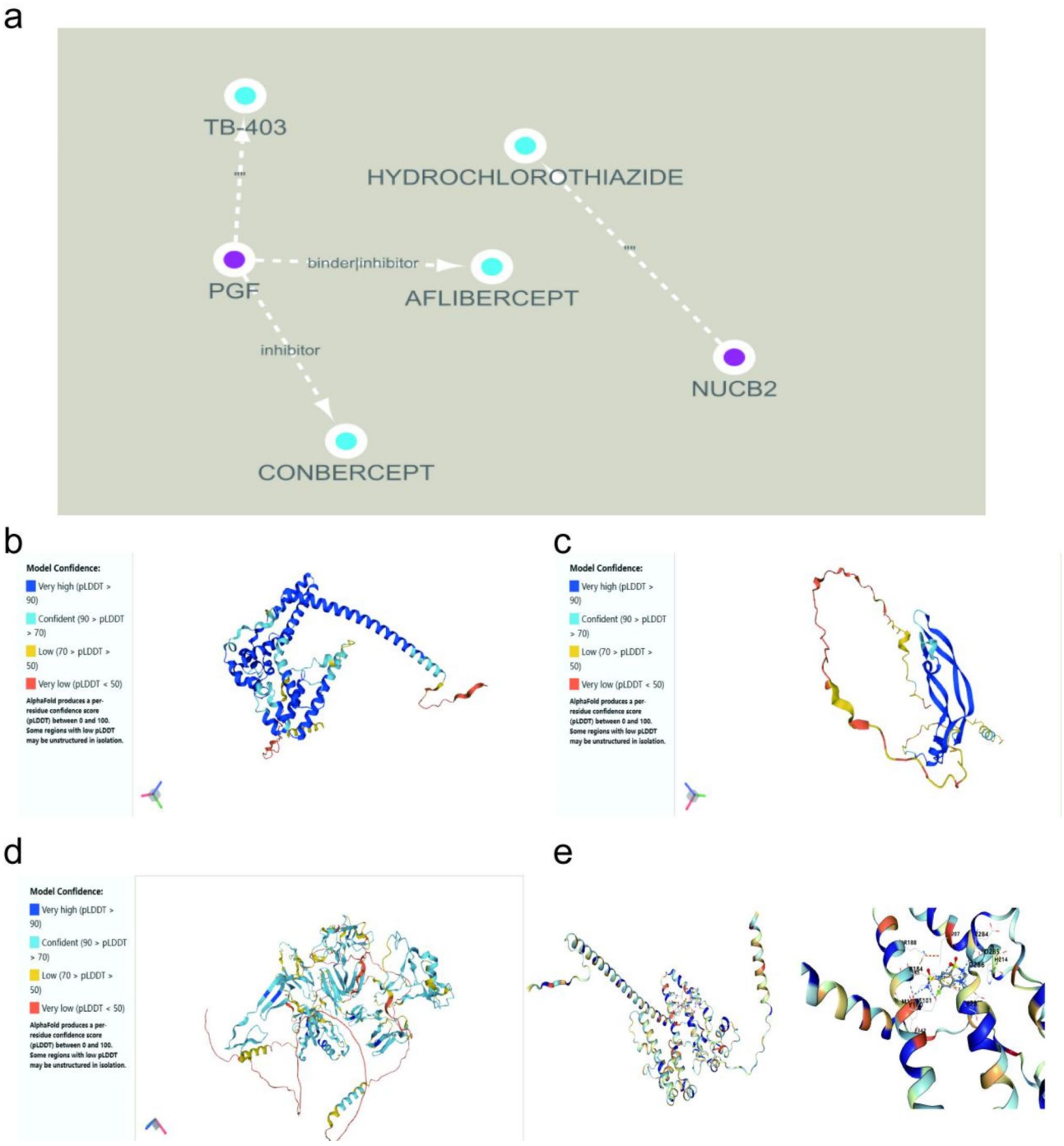

Fig. 6 Potential drug targets in key genes. (**a**) Drug prediction analysis for NUCB2, PGF, and CRIM1 genes. (**b**) 3D structure prediction model for the protein encoded by the NUCB2 gene. (**c**) 3D structure prediction model for the protein encoded by the PGF gene. (**d**) 3D structure prediction model for the protein encoded by the CRIM1 gene. (**e**) 3D structure prediction model for the protein encoded by the NUCB2 and HYDROCHLOROTHIAZIDE

and 7 being relatively scarce. We also showcased the trajectory graphs of fibroblasts in different samples, which clearly demonstrated that PCOS samples exhibited a higher abundance of fibroblasts (Fig. 8c). This might indicate enhanced tissue remodeling activity within the ovaries, potentially leading to cyst development and other structural changes associated with PCOS. Fibroblasts could contribute to characteristic manifestations of PCOS, such as ovarian fibrosis, altered folliculogenesis, and chronic inflammation. Furthermore, we visualized the expression changes of key genes along the developmental trajectory of fibroblasts (Fig. 8d). Observations

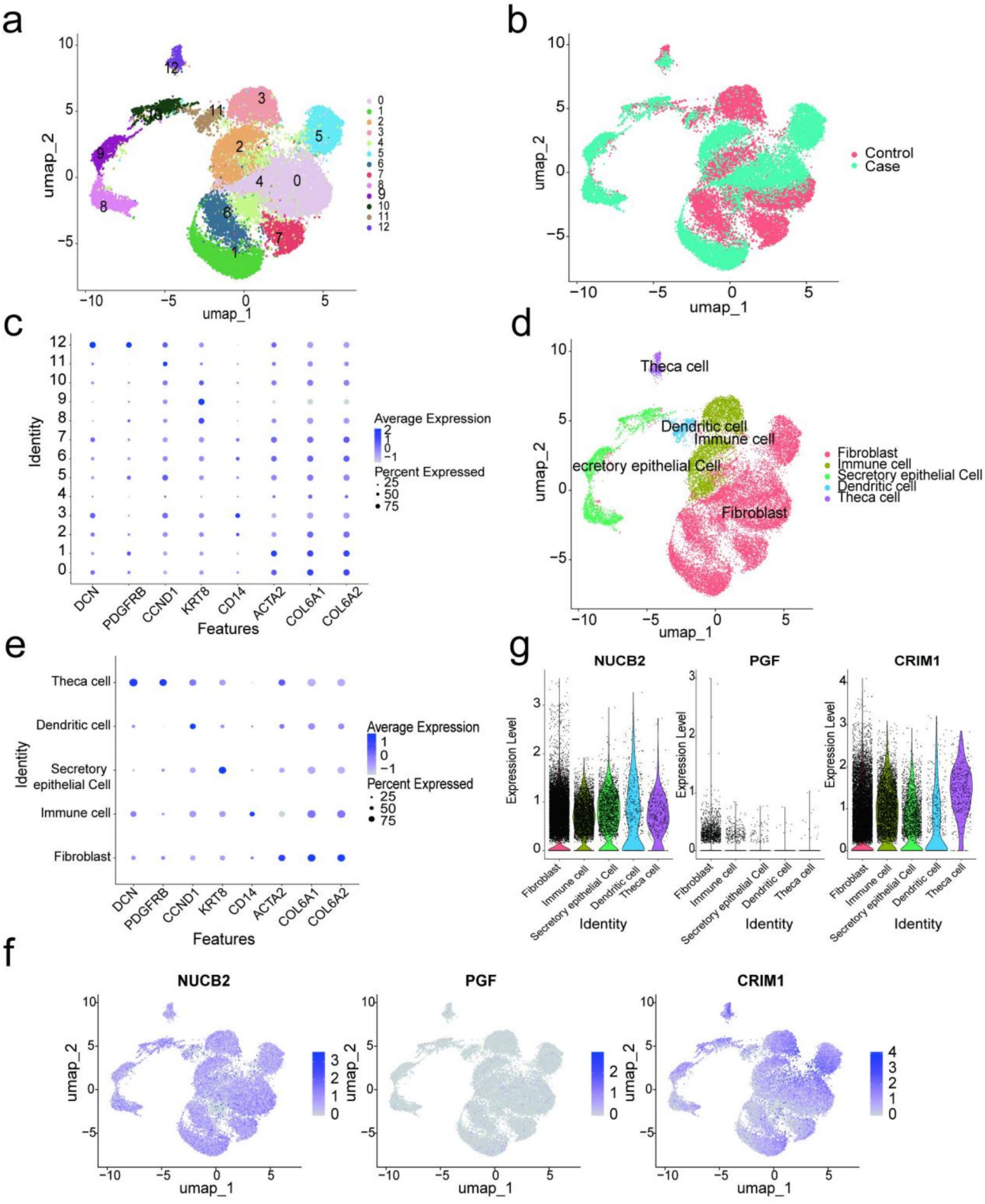

Fig. 7 Identification of key cell. (**a**) Quality control (QC) process of key cells. (**b**) Identified Cell Clusters Along Pseudotime Trajectories in Biological Processes.(**c**) Gene expression levels. (**d**) Annotation of five cell clusters. (**e**) Bubble plot of related marker genes. (**f**) Analysis of key gene expression in cell clusters. (**g**) Expression levels of key genes across five cellular clusters

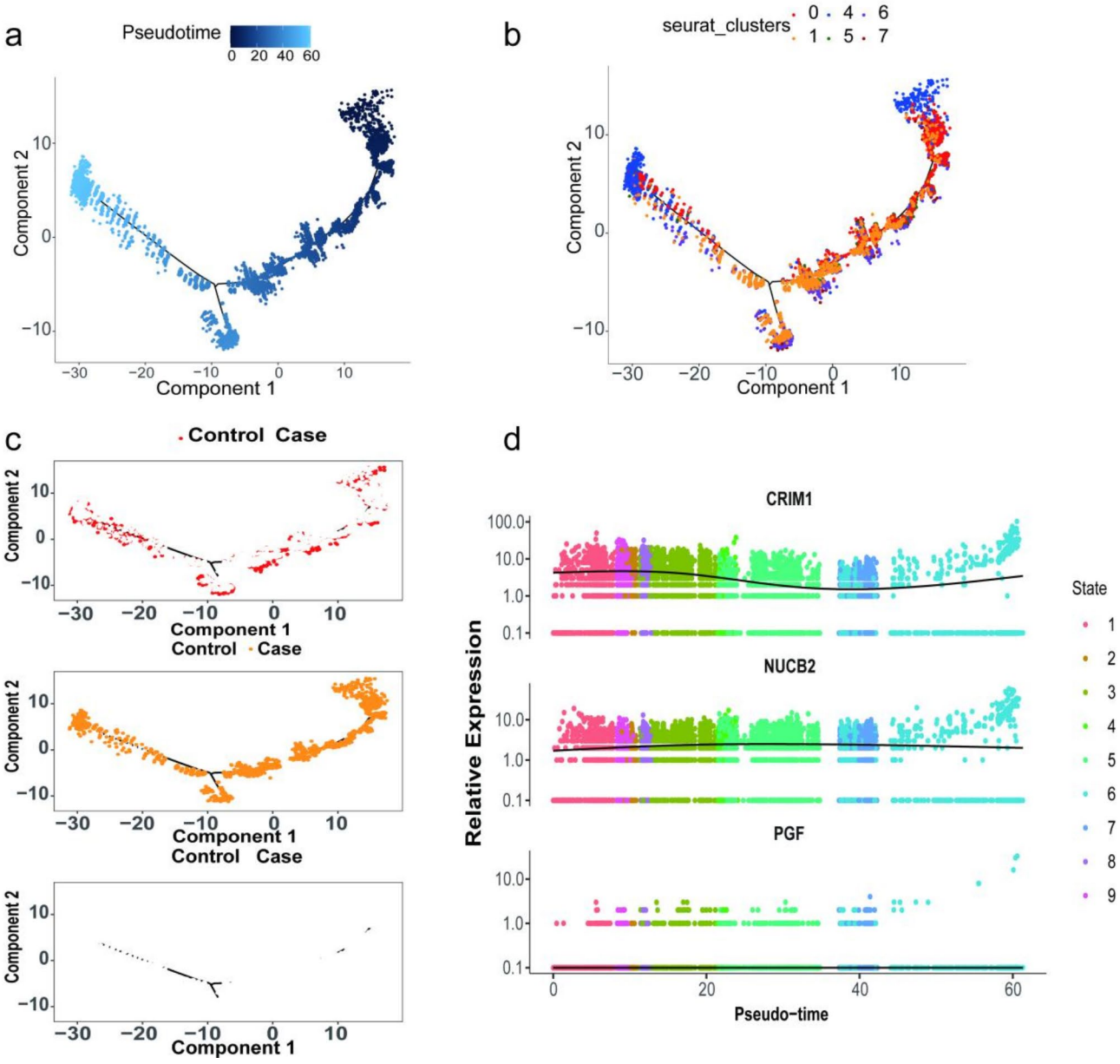

Fig. 8 Trajectories of fibroblast maturation and heterogeneity. (**a**) Pseudo-Time trajectory inference of fibroblasts. (**b**) Pseudo-Time trajectory inference analysis of six fibroblast cellular clusters. (**c**) The trajectory graphs of fibroblasts in different samples. (**d**) Expression dynamics of key genes during fibroblast development

revealed that as fibroblasts matured, NUCB2 expression initially increased then decreased, CRIM1 expression showed an initial increase followed by a decrease and then an increase again, while, consistent with previous findings identifying key cells, PGF expression remained low and stable throughout.

Communication analysis of cell clusters

To shed light on the intercellular interactions between different cell clusters in both PCOS and control samples, we delved into analyzing the cell communication network. It's evident that there existed a level of communication among the five identified cell clusters (Fig. 9a-d). To be specific, in the PCOS samples, the quantity and intensity of communication between secretory epithelial cells and fibroblasts were heightened, whereas in the control samples, it's the fibroblasts and immune cells that showcased a greater extent of communication both in number and significance. Moreover, the analysis revealed that in PCOS samples, the number of ligand-receptor interaction words between fibroblasts and other cells falls short compared to that in control samples. This diminished strength in cell-to-cell communication might be intricately linked to the myriad of

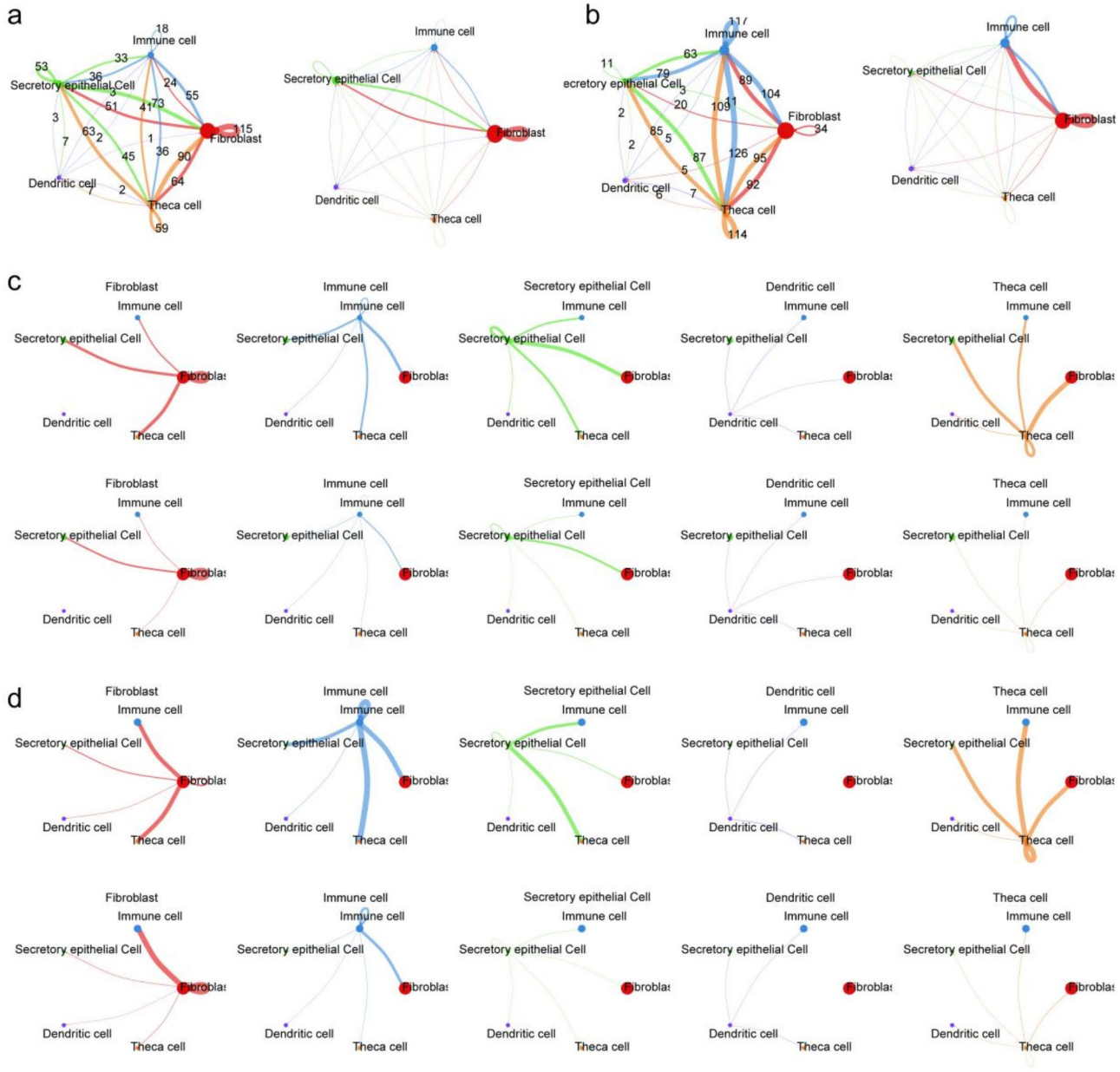

Fig. 9 Communication networks among cellular clusters. (**a**) Communication networks between secretory epithelial cell clusters and fibroblast cell clusters.(**b**) Communication networks between Immune cell clusters and fibroblast cell clusters. (**c**)Communication networks of a single cell cluster with four other cell clusters.(**d**)Ligand-receptor interactions between fibroblasts and other cell types

pathological characteristics attributed to PCOS, potentially compromising the crucial roles of fibroblasts in processes such as ovarian tissue remodeling, inflammatory responses, and follicle development. Consequently, this could lead to the manifestation of reproductive disorders and hormonal imbalances observed in individuals with PCOS.

Discussion

PCOS is a common endocrine disorder affecting female reproduction, with its specific etiology and pathogenesis remaining unclear [49]. EMT, the process in which epithelial cells transition to a mesenchymal phenotype, is closely linked to PCOS, although the specific mechanisms underlying this relationship remain unclear [8]. This study utilized transcriptome and scRNA-seq datasets to identify DEGs and DE-EMT-RGs in PCOS, conducting enrichment analyses and employing MR to explore causal relationships. The findings revealed that

key genes NUCB2, PGF, and CRIM1 are protective factors against PCOS and showed specific expression patterns across immune cells and organ tissues. Aflibercept and Conbercept were identified as potential inhibitors of PGF. scRNA-seq analysis showed that fibroblasts are key cells in PCOS. These results provide insights into potential therapeutic targets and molecular mechanisms.

Currently, there are few reports on the role of NUCB2 in the pathogenesis of PCOS. Nucleobindin-2, encoded by NUCB2, is a multifunctional calcium-binding protein that plays a role in various physiological and pathological processes, such as regulating insulin secretion and being involved in neurodegenerative diseases [50]. Our results showed that NUCB2 was significantly downregulated in granulosa cells of PCOS patients, and its role as a protective factor for PCOS was confirmed by MR analysis. Similarly, a study found that the expression level of NUCB2 was significantly reduced in the obese subtype of PCOS patients, suggesting that it may play a role in PCOS-related metabolic abnormalities [50]. Another study found that the NUCB2 knockout mouse model exhibited insulin resistance, hyperglycemia, and obesity phenotypes, which are clinical features of PCOS [51], further supporting the possibility that NUCB2 may influence the development of PCOS by regulating insulin sensitivity.

Placental Growth Factor (PGF), an angiogenic factor belonging to the VEGF family, is mainly involved in the regulation of angiogenesis and the promotion of trophoblast proliferation and migration. Currently, there are few reports on the role of PGF in the pathogenesis of PCOS. Our results showed that PGF was significantly upregulated in the granulosa cells of PCOS patients. Similarly, one study found that PGF was highly expressed in the ovarian tissues of PCOS patients, suggesting that it may be associated with hyperandrogenemia and elevated gonadotropin levels in PCOS [52].

In this study, we discovered for the first time that CRIM1 is significantly upregulated in PCOS and acts as a protective factor against the disease. CRIM1 is an important extracellular matrix (ECM) regulator that affects cell proliferation, differentiation, and migration, mainly through interaction with the bone morphogenetic protein (BMP) signaling pathway. Previous studies have demonstrated that CRIM1 plays a critical role in regulating placental development, organogenesis, angiogenesis, and is implicated in kidney disease and cancer [53].

The pathway analysis through three key genes mainly involved proteasome pathway, glycolysis and gluconeogenesis pathway, the ribosome pathway, and the axon guidance pathway. All pathways identified were important to PCOS. Multiple studies have identified significant abnormalities in the glycolysis and gluconeogenesis pathways in PCOS. Multiomics analysis of granulosa cells revealed differential gene expression within these metabolic pathways, highlighting their critical regulatory roles in steroid biosynthesis and metabolic signaling [26]. Concurrently, research on the metabolic characteristics of hepatic exosomes in PCOS mice demonstrated distinct age-related differences in substrate utilization for gluconeogenesis, suggesting that these metabolic changes are closely linked to insulin resistance and the progression of PCOS [54]. Studies have found that SYVN1 promotes the degradation of Drp1 in granulosa cells through the proteasome-dependent pathway, thereby inhibiting apoptosis and mitochondrial fission. This highlights the function of SYVN1 in polycystic ovary syndrome and provides insights into potential clinical treatment target [55]. A study that sequenced plasma exosomal miRNAs in PCOS patients identified five key miRNAs (miR-126-3p, miR-146a-5p, miR-20b-5p, miR-106a-5p, and miR-18a-3p). These miRNAs are involved in axon guidance and are related to the menstrual cycle, prefollicle numbers, and hormone levels [56].

Recent studies have highlighted the upregulation of critical ribosomal proteins such as Rps21 and Rpl36 in the oocytes of PCOS mice. This upregulation indicates a disruption in the protein synthesis machinery, essential for proper follicle development, thereby potentially contributing to PCOS pathophysiology by impairing oocyte quality and maturation [57]. Additionally, research has revealed that the ribosome pathway in PCOS is modulated by complex transcriptional regulatory networks involving specific transcription factors and miRNAs, altering ribosomal gene expression. This dysregulation impacts key cellular processes, contributing to the metabolic and reproductive abnormalities characteristic of PCOS, such as insulin resistance, increased androgen production, and disrupted folliculogenesis [58]. These findings underscore the significant role of ribosomal gene regulation in PCOS pathogenesis and suggest that targeting these pathways could offer promising therapeutic interventions.

NUCB2 and CRIM1 are highly expressed in plasmacytoid dendritic cells. Furthermore, the three key genes are co-expressed in smooth muscle.Their gene interactions suggest a role in regulating cell proliferation. Recent studies have identified a significant role for plasmacytoid dendritic cells (pDCs) in PCOS, particularly regarding insulin resistance, follicle development, and hyperandrogenism. pDCs exacerbate insulin resistance in PCOS by producing pro-inflammatory cytokines such as IFN-α, disrupting insulin signaling pathways [59]. This inflammatory milieu impairs folliculogenesis, contributing to the anovulation characteristic of PCOS [59]. Additionally, hyperandrogenism alters pDC activity, increasing inflammatory cytokine production and creating a feedback loop that intensifies both inflammation and androgen excess [60]. Targeting pDCs may, therefore, offer

novel therapeutic strategies to reduce inflammation and improve metabolic and reproductive outcomes in PCOS.

The construction of molecular regulatory networks for key genes NUCB2, PGF, and CRIM1 has revealed intricate interactions involving miRNAs, lncRNAs, and transcription factors. Specifically, a network consisting of 51 miRNAs and 17 lncRNAs was identified. Studies have demonstrated that exosomal miR-30c-5p derived from adipocytes activates signal transduction pathways in human ovarian microvascular endothelial cells. Consequently, this leads to the promotion of ovarian angiogenesis through the upregulation of the STAT3/VEGFA pathway by targeting SOCS3, thus contributing to the onset of symptoms associated with polycystic ovary syndrome [61]. Furthermore, investigations have revealed that the p-JAK2/p-STAT3 signaling cascade plays a regulatory role in follicular development in rodent models of polycystic ovary syndrome [62].

Drug prediction analysis has identified Aflibercept and Conbercept as inhibitors targeting PGF, a key gene in PCOS. A systematic review has demonstrated that Aflibercept is an effective treatment for age-related macular degeneration [63]. Animal studies show that Aflibercept significantly inhibits PGF and vascular endothelial growth factor-A, preventing vascular leakage and choroidal neovascularization [64]. Furthermore, PGF levels in the follicular fluid of women with PCOS were found to be 1.5 times higher, with bioavailability significantly increased by 2-fold compared to non-PCOS controls [52]. Aflibercept may inhibit PGF's role in promoting endothelial and epithelial cell proliferation, potentially mitigating the angiogenesis and inflammation characteristic of PCOS.In our molecular docking analysis, we have, for the first time, demonstrated the 3D structure of the interaction between hydrochlorothiazide and NUCB2. MR analysis suggests that NUCB2 acts as a protective factor in PCOS. This finding indicates that hydrochlorothiazide may exert therapeutic effects on PCOS through NUCB2 and its related pathways.

Our comprehensive single-cell, pseudotime, and cell communication analyses in PCOS identified fibroblasts as key cells showing distinct gene expression profiles with higher expressions of NUCB2, PGF, and CRIM1. Pseudotime trajectories indicated significant cellular differentiation and heterogeneity, suggesting an enhanced tissue remodeling activity that could contribute to ovarian fibrosis and altered folliculogenesis. Furthermore, altered cell communication patterns were observed, with increased interactions between secretory epithelial cells and fibroblasts in PCOS, potentially influencing ovarian tissue remodeling and inflammatory responses. Recent research highlights the significant role of fibroblasts in PCOS pathophysiology, particularly in insulin resistance, hyperandrogenism, and follicle development. Studies show that fibroblasts from PCOS patients have altered insulin receptor signaling, characterized by increased serine phosphorylation and decreased insulin-stimulated autophosphorylation, contributing to insulin resistance [65, 66]. Additionally, TGF-β modulates the expression of PCOS candidate genes in fetal ovarian fibroblasts, suggesting a mechanism through which fibroblasts influence ovarian function and PCOS development [67]. Furthermore, endometrial stromal fibroblasts in PCOS exhibit impaired progesterone responses, leading to aberrant decidualization and altered cytokine profiles, potentially affecting implantation and elevating the risk of endometrial pathologies [68]. These findings underscore the critical involvement of fibroblasts in the molecular and cellular mechanisms underlying PCOS, providing insights into potential therapeutic targets.

Our study has identified NUCB2, PGF, and CRIM1 as key genes in PCOS. Mendelian Randomization analysis has confirmed these genes as protective factors. Single-cell analysis highlighted fibroblasts as pivotal cells significantly involved in tissue remodeling and inflammation in PCOS. While our findings provide valuable insights, further experimental validation, and clinical studies are needed to confirm these mechanisms and evaluate the therapeutic potential of the predicted drugs.

Conclusion

Our study has significantly advanced the understanding of PCOS by elucidating the relationship between PCOS and epithelial-mesenchymal transition (EMT). We identified NUCB2, PGF, and CRIM1 as key genes involved in crucial pathways such as glycolysis, gluconeogenesis, and the proteasome pathway. Mendelian Randomization analysis confirmed these genes as protective factors, suggesting their potential as therapeutic targets. Single-cell analysis highlighted the pivotal role of fibroblasts in tissue remodeling and inflammation in PCOS. Importantly, our findings underscore the significant involvement of these genes in EMT-related processes, linking EMT mechanisms directly to PCOS pathophysiology. This connection opens new avenues for therapeutic strategies targeting EMT to manage PCOS more effectively. Future research should focus on experimental validation of these findings and the development of clinical applications to leverage EMT mechanisms for novel PCOS treatments.

Abbreviations

GEO	Gene Expression Omnibus
PCOS	Polycystic Ovary Syndrome
scRNA-seq	Ingle-cell RNA sequencing
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
BP	Biological Process
CC	Cellular Component
MF	Molecular Function
PPI	Protein-Protein Interaction
MR	Mendelian Randomization

eQTL	The Expression Quantitative Trait Loci
GWAS	Genome-wide Association Study
SNPs	Single nucleotide polymorphisms
LD	linkage disequilibrium
IVW	Inverse variance weighted
OR	Odds Ratio
LOO	Leave-One-Out
GSEA	Gene Set Enrichment Analysis
GSVA	Gene Set Variation Analysis
TFs	Transcription Factors
LDDT	The Local Distance Difference Test
QC	Quality Control
PC	Sprincipal Components
PCA	Principal Component Analysis
UMAP	Uniform Manifold Approximation and Projection

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13048-025-01617-2.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Acknowledgements
We would like to express our sincere gratitude to all individuals and organizations who supported and assisted us throughout this research.

Author contributions
Dong L.: Data curation, methodology, formal analysis, software, visualization, writing-original draft, writing-review&editing. Dan L.: Data curation, writing-original draft. K.Z.: Conceptualization, methodology, project administration, resources, supervision, writing-original draft, and writing-review&editing.

Funding
This study was supported by the Sichuan Provincial Natural Science Foundation Project (2025ZNSFSC1669) and the Key R&D Plan of the Sichuan Provincial Department of Science and Technology (No. 2023YFS0072).

Data availability
The dataset(s) supporting the conclusions of this article is(are) available in the Gene Expression Omnibus (GEO) database repository, [unique persistent identifier and hyperlink to dataset(s) in https://www.ncbi.nlm.nih.gov/gds].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Obstetrics and Gynecology, West China Second University Hospital, Sichuan University, Chengdu, China
[2]Key Laboratory of Birth Defects and Related Diseases of Women and Children (Sichuan University), Ministry of Education, Chengdu, China
[3]Department of Ultrasonic Medicine, West China Second University Hospital of Sichuan University, Chengdu, China

Received: 22 August 2024 / Accepted: 4 February 2025
Published online: 19 February 2025

References

1. Escobar-Morreale HF. Polycystic ovary syndrome: definition, aetiology, diagnosis and treatment. Nat Rev Endocrinol. 2018;14 5:270–84. https://doi.org/10.1038/nrendo.2018.24.
2. Szczuko M, Kikut J, Szczuko U, Szydlowska I, Nawrocka-Rutkowska J, Zietek M, et al. Nutrition Strategy and Life Style in Polycystic Ovary syndrome-narrative review. Nutrients. 2021;13:7. https://doi.org/10.3390/nu13072452.
3. Balen AH, Morley LC, Misso M, Franks S, Legro RS, Wijeyaratne CN, et al. The management of anovulatory infertility in women with polycystic ovary syndrome: an analysis of the evidence to support the development of global WHO guidance. Hum Reprod Update. 2016;22 6:687–708. https://doi.org/10.1093/humupd/dmw025.
4. Al Wattar BH, Fisher M, Bevington L, Talaulikar V, Davies M, Conway G, et al. Clinical practice guidelines on the diagnosis and management of polycystic ovary syndrome: a systematic review and Quality Assessment Study. J Clin Endocrinol Metab. 2021;106 8:2436–46. https://doi.org/10.1210/clinem/dgab232.
5. Sadeghi HM, Adeli I, Calina D, Docea AO, Mousavi T, Daniali M, et al. Polycystic ovary syndrome: a Comprehensive Review of Pathogenesis, Management, and Drug Repurposing. Int J Mol Sci. 2022;23(2). https://doi.org/10.3390/ijms23020583.
6. Chen W, Yang Q, Hu L, Wang M, Yang Z, Zeng X, et al. Shared diagnostic genes and potential mechanism between PCOS and recurrent implantation failure revealed by integrated transcriptomic analysis and machine learning. Front Immunol. 2023;14:1175384. https://doi.org/10.3389/fimmu.2023.1175384.
7. Marconi GD, Fonticoli L, Rajan TS, Pierdomenico SD, Trubiani O, Pizzicannella J, et al. Epithelial-mesenchymal transition (EMT): the Type-2 EMT in Wound Healing, tissue regeneration and Organ Fibrosis. Cells. 2021;10:7. https://doi.org/10.3390/cells10071587.
8. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. J Clin Invest. 2009;119 6:1420–8. https://doi.org/10.1172/JCI39104.
9. Sarrand J, Soyfoo MS. Involvement of epithelial-mesenchymal transition (EMT) in Autoimmune diseases. Int J Mol Sci. 2023;24:19. https://doi.org/10.3390/ijms241914481.
10. Zheng Q, Liu M, Fu J. ALG2 inhibits the epithelial-to-mesenchymal transition and stemness of ovarian granulosa cells through the Wnt/beta-catenin signaling pathway in polycystic ovary syndrome. Reprod Biol. 2022;22 4:100706. https://doi.org/10.1016/j.repbio.2022.100706.
11. Makieva S, Giacomini E, Ottolina J, Sanchez AM, Papaleo E, Vigano P. Inside the Endometrial Cell Signaling Subway: mind the gap(s). Int J Mol Sci. 2018;19:9. https://doi.org/10.3390/ijms19092477.
12. Hu M, Zhang Y, Li X, Cui P, Li J, Brannstrom M, et al. Alterations of endometrial epithelial-mesenchymal transition and MAPK signalling components in women with PCOS are partially modulated by metformin in vitro. Mol Hum Reprod. 2020;26 5:312–26. https://doi.org/10.1093/molehr/gaaa023.
13. Wang Y, Leung P, Li R, Wu Y, Huang H, Editorial. Polycystic ovary syndrome (PCOS): mechanism and management. Front Endocrinol (Lausanne). 2022;13:1030353. https://doi.org/10.3389/fendo.2022.1030353.
14. Sekula P, Del Greco MF, Pattaro C, Kottgen A. Mendelian randomization as an Approach to assess causality using Observational Data. J Am Soc Nephrol. 2016;27 11:3253–65. https://doi.org/10.1681/ASN.2016010098.
15. Bowden J, Holmes MV. Meta-analysis and mendelian randomization: a review. Res Synth Methods. 2019;10 4:486–96. https://doi.org/10.1002/jrsm.1346.
16. Li JW, Chen YZ, Zhang Y, Zeng LH, Li KW, Xie BZ, et al. Gut microbiota and risk of polycystic ovary syndrome: insights from mendelian randomization. Heliyon. 2023;9 12:e22155. https://doi.org/10.1016/j.heliyon.2023.e22155.
17. Cheng G, Wang M, Sun H, Lai J, Feng Y, Liu H, et al. Age at menopause is inversely related to the prevalence of common gynecologic cancers: a study based on NHANES. Front Endocrinol (Lausanne). 2023;14:1218045. https://doi.org/10.3389/fendo.2023.1218045.
18. Aru N, Yang C, Chen Y, Liu J. Causal association of immune cells and polycystic ovarian syndrome: a mendelian randomization study. Front Endocrinol (Lausanne). 2023;14:1326344. https://doi.org/10.3389/fendo.2023.1326344.
19. Zhu T, Goodarzi MO. Causes and consequences of polycystic ovary syndrome: insights from mendelian randomization. J Clin Endocrinol Metab. 2022;107 3:e899–911. https://doi.org/10.1210/clinem/dgab757.
20. Du Y, Li F, Li S, Ding L, Liu M. Causal relationship between polycystic ovary syndrome and chronic kidney disease: a mendelian randomization study.

Front Endocrinol (Lausanne). 2023;14:1120119. https://doi.org/10.3389/fendo.2023.1120119.
21. Wang Y, Wang JY, Schnieke A, Fischer K. Advances in single-cell sequencing: insights from organ transplantation. Mil Med Res. 2021;8(1:45). https://doi.org/10.1186/s40779-021-00336-1.
22. Wang S, Sun ST, Zhang XY, Ding HR, Yuan Y, He JJ, et al. The evolution of single-cell RNA sequencing technology and application: progress and perspectives. Int J Mol Sci. 2023;24(3). https://doi.org/10.3390/ijms24032943.
23. Harris RA, McAllister JM, Strauss JF. 3rd. Single-cell RNA-Seq identifies pathways and genes contributing to the Hyperandrogenemia Associated with Polycystic Ovary Syndrome. Int J Mol Sci. 2023;24:13. https://doi.org/10.3390/ijms241310611.
24. Qi L, Liu B, Chen X, Liu Q, Li W, Lv B, et al. Single-Cell Transcriptomic Analysis Reveals Mitochondrial Dynamics in oocytes of patients with polycystic ovary syndrome. Front Genet. 2020;11:396. https://doi.org/10.3389/fgene.2020.00396.
25. Li J, Chen H, Gou M, Tian C, Wang H, Song X, et al. Molecular features of polycystic ovary syndrome revealed by Transcriptome Analysis of Oocytes and Cumulus cells. Front Cell Dev Biol. 2021;9:735684. https://doi.org/10.3389/fcell.2021.735684.
26. Zhao R, Jiang Y, Zhao S, Zhao H. Multiomics Analysis Reveals Molecular Abnormalities in Granulosa cells of women with polycystic ovary syndrome. Front Genet. 2021;12:648701. https://doi.org/10.3389/fgene.2021.648701.
27. Love MI, Huber W, Anders S. Moderated estimation of Fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15 12:550. https://doi.org/10.1186/s13059-014-0550-8.
28. Gustavsson EK, Zhang D, Reynolds RH, Garcia-Ruiz S, Ryten M. Ggtranscript: an R package for the visualization and interpretation of transcript isoforms using ggplot2. Bioinformatics. 2022;38 15:3844–6. https://doi.org/10.1093/bioinformatics/btac409.
29. Gu Z, Hubschmann D. Make interactive Complex heatmaps in R. Bioinformatics. 2022;38 5:1460–2. https://doi.org/10.1093/bioinformatics/btab806.
30. Zheng Y, Gao W, Zhang Q, Cheng X, Liu Y, Qi Z, et al. Ferroptosis and autophagy-related genes in the pathogenesis of ischemic cardiomyopathy. Front Cardiovasc Med. 2022;9:906753. https://doi.org/10.3389/fcvm.2022.906753.
31. 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. https://doi.org/10.1089/omi.2011.0118.
32. Liu P, Xu H, Shi Y, Deng L, Chen X. Potential molecular mechanisms of Plantain in the treatment of gout and hyperuricemia based on Network Pharmacology. Evid Based Complement Alternat Med. 2020;2020:3023127. https://doi.org/10.1155/2020/3023127.
33. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7. https://doi.org/10.7554/eLife.34408.
34. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015;44(2):512–25. https://doi.org/10.1093/ije/dyv080.
35. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in mendelian randomization with some Invalid instruments using a weighted median estimator. Genet Epidemiol. 2016;40 4:304–14. https://doi.org/10.1002/gepi.21965.
36. Burgess S, Scott RA, Timpson NJ, Davey Smith G, Thompson SG, Consortium E-I. Using published data in mendelian randomization: a blueprint for efficient identification of causal risk factors. Eur J Epidemiol. 2015;30 7:543–52. https://doi.org/10.1007/s10654-015-0011-z.
37. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. Int J Epidemiol. 2017;46 6:1985–98. https://doi.org/10.1093/ije/dyx102.
38. Qin Q, Zhao L, Ren A, Li W, Ma R, Peng Q, et al. Systemic lupus erythematosus is causally associated with hypothyroidism, but not hyperthyroidism: a mendelian randomization study. Front Immunol. 2023;14:1125415. https://doi.org/10.3389/fimmu.2023.1125415.
39. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014;23 R1:R89–98. https://doi.org/10.1093/hmg/ddu328.
40. Cui Z, Feng H, He B, He J, Tian Y. Relationship between serum amino acid levels and bone Mineral density: a mendelian randomization study. Front Endocrinol (Lausanne). 2021;12:763538. https://doi.org/10.3389/fendo.2021.763538.
41. Xiao G, He Q, Liu L, Zhang T, Zhou M, Li X, et al. Causality of genetically determined metabolites on anxiety disorders: a two-sample mendelian randomization study. J Transl Med. 2022;20(1:475). https://doi.org/10.1186/s12967-022-03691-2.
42. Correction to Lancet Psych. (22)00377-7. Lancet Psychiatry. 2023;10 2:e5. https://doi.org/10.1016/S2215-0366. 2022 published online Dec 8.
43. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7. https://doi.org/10.1186/1471-2105-14-7.
44. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43 7:e47. https://doi.org/10.1093/nar/gkv007.
45. Zhang H, Meltzer P, Davis S. RCircos: an R package for Circos 2D track plots. BMC Bioinformatics. 2013;14:244. https://doi.org/10.1186/1471-2105-14-244.
46. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33 5:495–502. https://doi.org/10.1038/nbt.3192.
47. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32 4:381–6. https://doi.org/10.1038/nbt.2859.
48. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, et al. Inference and analysis of cell-cell communication using CellChat. Nat Commun. 2021;12(1:1088). https://doi.org/10.1038/s41467-021-21246-9.
49. Lizneva D, Suturina L, Walker W, Brakta S, Gavrilova-Jordan L, Azziz R. Criteria, prevalence, and phenotypes of polycystic ovary syndrome. Fertil Steril. 2016;106 1:6–15. https://doi.org/10.1016/j.fertnstert.2016.05.003.
50. Taskin MI, Eser B, Adali E, Kara H, Cuce C, Hismiogullari AA. NUCB2 gene polymorphism and its relationship with nesfatin-1 levels in polycystic ovary syndrome. Gynecol Endocrinol. 2016;32 1:46–50. https://doi.org/10.3109/09513590.2015.1081682.
51. Gharanei S, Ramanjaneya M, Patel AH, Patel V, Shabir K, Auld C, et al. NUCB2/Nesfatin-1 reduces obesogenic Diet Induced inflammation in mice Subcutaneous White Adipose tissue. Nutrients. 2022;14:7. https://doi.org/10.3390/nu14071409.
52. Tal R, Seifer DB, Grazi RV, Malter HE. Follicular fluid placental growth factor is increased in polycystic ovarian syndrome: correlation with ovarian stimulation. Reprod Biol Endocrinol. 2014;12:82. https://doi.org/10.1186/1477-7827-12-82.
53. Zeng H, Tang L. CRIM1, the antagonist of BMPs, is a potential risk factor of cancer. Curr Cancer Drug Targets. 2014;14 7:652–8. https://doi.org/10.2174/1568009614666140725094125.
54. Gao S, Long F, Jiang Z, Shi J, Ma D, Yang Y, et al. The complex metabolic interactions of liver tissue and hepatic exosome in PCOS mice at young and middle age. Front Physiol. 2022;13:990987. https://doi.org/10.3389/fphys.2022.990987.
55. Sun L, Ye H, Tian H, Xu L, Cai J, Zhang C, et al. The E3 ubiquitin ligase SYVN1 plays an antiapoptotic role in polycystic ovary syndrome by regulating mitochondrial fission. Oxid Med Cell Longev. 2022;2022:3639302. https://doi.org/10.1155/2022/3639302.
56. Jiang X, Li J, Zhang B, Hu J, Ma J, Cui L, et al. Differential expression profile of plasma exosomal microRNAs in women with polycystic ovary syndrome. Fertil Steril. 2021;115 3:782–92. https://doi.org/10.1016/j.fertnstert.2020.08.019.
57. Miao C, Chen Y, Fang X, Zhao Y, Wang R, Zhang Q. Identification of the shared gene signatures and pathways between polycystic ovary syndrome and endometrial cancer: an omics data based combined approach. PLoS ONE. 2022;17 7:e0271380. https://doi.org/10.1371/journal.pone.0271380.
58. Sikiru AB, Adeniran MA, Akinola K, Behera H, Kalaignazhal G, Egena SSA-JMEFSJ. Unraveling the complexity of the molecular pathways associated with polycystic ovary syndrome (PCOS) and identifying molecular targets for therapeutic development: a review of literature. 2023;28 1:16.
59. Armanini D, Boscaro M, Bordin L, Sabbadin C. Controversies in the Pathogenesis, diagnosis and treatment of PCOS: focus on insulin resistance, inflammation, and Hyperandrogenism. Int J Mol Sci. 2022;23:8. https://doi.org/10.3390/ijms23084110.
60. Shabbir S, Khurram E, Moorthi VS, Eissa YTH, Kamal MA, Butler AE. The interplay between androgens and the immune response in polycystic ovary syndrome. J Transl Med. 2023;21(1:259). https://doi.org/10.1186/s12967-023-04116-4.
61. Hu J, Lin F, Yin Y, Shang Y, Xiao Z, Xu W. Adipocyte-derived exosomal miR-30c-5p promotes ovarian angiogenesis in polycystic ovary syndrome via the

SOCS3/STAT3/VEGFA pathway. J Steroid Biochem Mol Biol. 2023;230:106278. https://doi.org/10.1016/j.jsbmb.2023.106278.
62. Wang H, Feng X, Wang T, Pan J, Zheng Z, Su Y, et al. Role and mechanism of the p-JAK2/p-STAT3 signaling pathway in follicular development in PCOS rats. Gen Comp Endocrinol. 2023;330:114138. https://doi.org/10.1016/j.ygcen.2022.114138.
63. Heier JS, Brown DM, Chong V, Korobelnik JF, Kaiser PK, Nguyen QD, et al. Intravitreal Aflibercept (VEGF trap-eye) in wet age-related macular degeneration. Ophthalmology. 2012;119 12:2537–48. https://doi.org/10.1016/j.ophtha.2012.09.006.
64. Balser C, Wolf A, Herb M, Langmann T. Co-inhibition of PGF and VEGF blocks their expression in mononuclear phagocytes and limits neovascularization and leakage in the murine retina. J Neuroinflammation. 2019;16(1:26). https://doi.org/10.1186/s12974-019-1419-2.
65. Li M, Youngren JF, Dunaif A, Goldfine ID, Maddux BA, Zhang BB, et al. Decreased insulin receptor (IR) autophosphorylation in fibroblasts from patients with PCOS: effects of serine kinase inhibitors and IR activators. J Clin Endocrinol Metab. 2002;87 9:4088–93. https://doi.org/10.1210/jc.2002-020363.
66. Dunaif A, Xia J, Book CB, Schenker E, Tang Z. Excessive insulin receptor serine phosphorylation in cultured fibroblasts and in skeletal muscle. A potential mechanism for insulin resistance in the polycystic ovary syndrome. J Clin Invest. 1995;96(2):801–10. https://doi.org/10.1172/JCI118126.
67. Azumah R, Liu M, Hummitzsch K, Bastian NA, Hartanti MD, Irving-Rodgers HF, et al. Candidate genes for polycystic ovary syndrome are regulated by TGFbeta in the bovine foetal ovary. Hum Reprod. 2022;37 6:1244–54. https://doi.org/10.1093/humrep/deac049.
68. Piltonen TT, Chen JC, Khatun M, Kangasniemi M, Liakka A, Spitzer T, et al. Endometrial stromal fibroblasts from women with polycystic ovary syndrome have impaired progesterone-mediated decidualization, aberrant cytokine profiles and promote enhanced immune cell migration in vitro. Hum Reprod. 2015;30 5:1203–15. https://doi.org/10.1093/humrep/dev055.

Publisher's note