Transcriptome-wide association study identified candidate genes associated with polycystic ovary syndrome

Yingzhou Shi¹,²,³# Changrong Li²,³#, Hongjie Yu²,³#, Xinyu Chen¹,³, Ping Shi¹,³, Chao Xu¹,²,³*

¹Department of Endocrinology, Shandong Provincial Hospital, Shandong University, Jinan, Shandong, 250021, China.

²Department of Endocrinology, Shandong Provincial Hospital affiliated to Shandong First Medical University, Jinan, Shandong, 250021, China.

³Shandong Provincial Key Laboratory of Endocrinology and Lipid Metabolism; Institute of Endocrinology and Metabolism, Shandong Academy of Clinical Medicine, Jinan, Shandong, 250021, China.

#These authors contributed equally to this work.

*Corresponding author: Chao Xu  Email: doctorxuchao@163.com

Keyword: Transcriptome-wide association study, polycystic ovary syndrome,

Abstract:
Polycystic ovary syndrome (PCOS) is one of the most common metabolism and endocrine disorder affecting women of reproductive age. However, the genetic mechanisms of PCOS remain largely elusive. A transcriptome-wide association study (TWAS) was conducted to identify the genes associated with PCOS using gene expression references of the whole blood and ovary. Then mRNA expression profiling analysis was performed to identify common genes with TWAS. Genes detected by TWAS were subjected to Gene Ontology (GO), pathway enrichment and protein-protein analysis. TWAS identified 422 genes for the whole blood and 177 genes for ovary, such as EML6 (TWAS.P = 6.33E-05), GRAP (TWAS.P = 0.000211), MAFTRR (TWAS.P = 1.21E-05). Functional enrichment analysis detected biological processes including ncRNA metabolic process, microtubule-related pathways. The present study identified a number of potential PCOS genes and biological processes. These results may provide insight into the genetic basis of PCOS.
Introduction:

Polycystic ovary syndrome (PCOS), characterized by hyperandrogenism, irregular anovulation, and polycystic ovarian changes, is one of the most common reproductive metabolic disorders affecting reproductive-aged women. It is estimated that PCOS affect 6–20% of women\(^1\), and approximately 80% of those PCOS are infertile\(^2\). Although studies have suggested that obesity, age at menopause, insulin resistance/hyperinsulinemia, sex hormone-binding globulin (SHBG) levels, and depression may be risk factors for PCOS, its potential etiologies remain uncertain\(^3\).

Genetic factors are thought to be important in the pathogenesis of PCOS. For example, researchers have found that PCOS is twice as common in monozygotic twins as in dizygotic twins and conclude that genetic factors account for about 79% of the liability of PCOS\(^4\). The prevalence of PCOS is also higher among the mothers and sisters of affected women, suggesting a phenomenon of familial aggregation\(^5\). Several genome-wide association studies (GWAS) have identified susceptibility loci for PCOS\(^6-10\). However, the biological significance of their findings remains largely unclear as close to 90% of GWAS-identified SNPs are in the noncoding region. Expression quantitative trait loci (eQTL) are genomic loci that are associated with gene transcript levels. Gene expression is a critical intermediate phenotype between genes and diseases. Such integration of GWAS and eQTL is effective in identifying candidate genes associated with diseases.

The availability and cost of specimens, as well as intrinsic factors, small effects, make large-scale expression-trait associations challenging. To address these issues, transcriptome-wide association studies (TWAS) were developed, which combine gene expression data with large-scale genome-wide association studies. First, datasets (e.g. GTEx) with matched expression data and genotype were used for building a prediction model. Then GWAS data are combined with these models to assess the association between predicted gene expression and disease risk. A growing number of researchers are using TWASs to identify disease-associated genetic loci\(^11-13\).

In the present study, we conduct TWAS and mRNA expression profiling analysis to identify genes that are associated with PCOS. Initially, TWAS was performed to identify genes associated with PCOS risk. Then mRNA expression profiling analysis was carried out to identify common genes between TWAS. For the genes identified by TWAS, we conducted an analysis of gene ontology and pathway enrichment to explore their biological functions and related pathways.
Method:

GWAS summary data for PCOS
We obtained PCOS GWAS summary data from a GWAS meta-analysis of seven cohort studies that included 10074 patients with PCOS and 103164 controls of European ancestry. The cohort studies include Rotterdam (cases: 1184, controls: 5799), UK (cases: 670, controls: 1379), EGCUT (cases: 157, controls: 2807), deCODE (cases: 658, controls: 6774), Chicago (cases: 984, controls: 2963), Boston (cases: 485, controls: 407), 23andMe (cases: 5184, controls: 82759). Diagnosis of PCOS was made based on National Institutes of Health criteria, Rotterdam criteria, or self-report.

TWAS analysis of PCOS
TWAS analysis was conducted using FUSION software. Pre-computed gene expression weights combined with PCOS summary statistics were used to calculate the association between single gene and PCOS. We derived the gene expression weights for whole blood and the ovary from the FUSION website (http://gusevlab.org/projects/fusion). In FUSION, expression weights were calculated using five linear models, including the best linear unbiased prediction model (BLUP), Bayesian sparse linear mixed model (BSLMM), least absolute shrinkage and selection operator (LASSO), Elastic Net, and top single nucleotide polymorphisms (SNPs). In the present study, a TWAS P-value was calculated for each gene within the whole blood and ovary for European populations.

mRNA Expression profiles of PCOS
A genome-wide mRNA expression profile of PCOS was used to identify differentially expressed genes (DEGs). The PCOS expression data were downloaded from the Gene Expression Omnibus (GEO) Datasets GSE34526 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE34526). Gonadotropin-stimulated granulosa cells were obtained from 7 patients with PCOS and 3 healthy individuals. GEO2R was used to screen for differentially expressed genes. Genes with a folding change (FC) >2 or <0.5 (|log2FC| > 1) and p-value < 0.01 were considered to be differentially expressed genes.

Functional Enrichment
Metascape (https://metascape.org/gp/index.html) was used to perform gene ontology and pathway enrichment analyses of the genes identified by TWAS. Metascape analysis is based on databases, including Gene Ontology, KEGG, Reactome Gene Sets, and WikiPathways. A hypergeometric test is used to identify enhanced ontology terms, and the p-value is corrected for false discovery rates. Molecular Complex Detection (MCODE) algorithm was used to extract densely connected regions.
Results:

TWAS results of PCOS

A total of genes were identified by TWAS in blood and ovary tissue.

The TWAS identified 422 genes with a p value < 0.05 for the whole blood, such as EML6 (TWAS.P = 6.33E-05), GRAP (TWAS.P = 0.000211), PTGER2 (TWAS.P = 0.000403) (Supplementary Table 1 and Table 2). For the ovary tissue, TWAS identified 177 genes with a p value < 0.05, such as MAFTRR (TWAS.P = 1.21E-05), NEIL2 (TWAS.P = 3.22E-05), ZNF555 (TWAS.P = 0.000364). 52 genes with p value < 0.05 were detected in both whole blood and ovary TWAS analysis (Supplementary Table 3). Table 1 and Figure 1 show the top 20 genes associated with PCOS as determined by TWAS.

Genes shared between TWAS and mRNA Expression Profiling

The TWAS analysis and the mRNA expression analysis detected 18 common genes, such as CAMK1D, COLEC12 and PLCB2 (Table 2). A Venn diagram of TWAS versus common genes identified by mRNA expression profiling are shown in Figure 2.

Gene Enrichment Analysis of genes identified by TWAS

In order to obtain an overall impression of the characteristics of these genes identified by TWAS, a total 547 genes were submitted to Metascape performing GO and pathway enrichment, and protein-protein interaction analysis.

GO and pathway enrichment analysis showed that TWAS-identified genes were enriched in ncRNA metabolic process, rRNA metabolic process, transport along microtubule and so on (Figure 3). MCODE enrichment analysis based on PPI enrichment for TWAS-identified genes showed 11 PPI modules (Figure 4). The top enrichment term of MCODE1 included 13 proteins distrusted in three major clusters: “Metabolism of RNA”, “Major pathway of rRNA processing in the nucleolus and cytosol”, “rRNA processing in the nucleus and cytosol”. MCODE2, which consists of 7 proteins was mostly enriched in “Interleukin-20 family signaling”, “JAK-STAT signaling pathway”, “Regulatory circuits of the STAT3 signaling pathway”.
Discussion
The TWAS method is a powerful tool that utilizes genetic variation and gene expression to identify genes whose cis-regulated expression is associated with complex traits. In present study, we performed TWAS to detect candidate genes closely associated with PCOS. Further integrative analysis of TWAS and mRNA expression profiling identified 18 common genes. To the best of our knowledge, this is the first TWAS integrated with mRNA expression profiling analysis for PCOS. The results of our study provide new insight into the genetic mechanism behind PCOS.

Our TWAS identified several genes that have been previously reported in GWAS and post-GWAS, such as RPS26\(^{19}\), NEIL2\(^{8}\), and ZNF556\(^{19}\) as well as novel genes that have never been reported.

NEIL2 is the gene encoding a member of the Fpg / Nei DNA glycosylase family. The encoded enzyme is involved in DNA repair mainly by cleaving oxidatively damaged bases and introducing DNA strand breaks through its base cleavage enzyme activity.

The NEIL2 gene encodes NEI endonuclease VIII-like 2, one of the DNA glycosylases involved in DNA repair\(^{20}\). DNA damage has been demonstrated to play a role in the pathogenesis of PCOS\(^{21}\). Further experiments are needed to determine the relationship between NEIL2 and PCOS.

EML6 is the only member of the EML family that is preferentially expressed by oocytes in the mouse ovary\(^{24}\). There is a colocalization of EML6 with meiotic spindles, and this specialized localization plays a critical role in maintaining spindle integrity and homologous chromosome segregation.

RPS26, located on 12q13.2, is a gene encoding a ribosomal protein, which is a component of the 40S subunit and belongs to the ribosomal protein S26E family\(^{22}\). The knockdown of Rps26 in mouse oocytes has previously been shown to result in delayed oocyte development and even the cessation of chromatin structural transition in oocytes, resulting in premature ovarian failure (POF)\(^{23}\).

GO and pathway enrichment indicated that ncRNA metabolic and microtubule-related pathways were enriched. The non-coding RNAs (ncRNAs) are functional RNAs that are not coded for protein production but serve as important regulators of a wide range of biological events\(^{25}\). Several studies have indicated that ncRNAs play a role in the occurrence and development of PCOS. The expression of snRNAs was significantly different in serum, granulosa cells (GCs), follicular fluid (FF), and other tissues between PCOS patients and the general population\(^{26,27}\). Accordingly, our results further confirm that ncRNA may play a crucial role in the development of PCOS.

Microtubules, which are polymers containing \(\alpha\)- and \(\beta\)-tubulin heterodimers, make up the oocyte spindle\(^{28}\). An abnormal spindle can affect the development, differentiation, and growth of a germ cell directly. Recent research has identified several mutations of spindle-related proteins that lead to infertility. An example of this is the TUBB8 gene, which encodes an isoform of beta-tubulin that is specific to humans and
primates. TUBB8 mutations in women lead to a range of spindle defects during oocyte meiosis. The enriched microtubule-related pathways may suggest that microtubules contribute to infertility in women with PCOS.

Conclusion:
In summary, this TWAS study identified novel and susceptible genes and potential pathways for PCOS. Further studies are needed to explore the potential mechanism.

Acknowledgements
We would like to thank the participants and the investigators for generating the publicly available dataset. The scientific calculations in this paper have been done on the HPC Cloud Platform of Shandong University.
Reference


Table 1 Top 20 genes identified by TWAS analysis

<table>
<thead>
<tr>
<th>Gene</th>
<th>Ensemble</th>
<th>Best GWAS ID</th>
<th>TWAS P-value</th>
<th>Tissue</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAFTRR</td>
<td>ENSG0000026139 0</td>
<td>rs1776738 3</td>
<td>0.0000121</td>
<td>Ovary</td>
</tr>
<tr>
<td>NEIL2</td>
<td>ENSG0000015432 8</td>
<td>rs2686206 8</td>
<td>0.0000322</td>
<td>Ovary</td>
</tr>
<tr>
<td>EML6</td>
<td>ENSG0000021459 5</td>
<td>rs1189874 7</td>
<td>0.0000633</td>
<td>Whole blood</td>
</tr>
<tr>
<td>GRAP</td>
<td>ENSG0000015401 6</td>
<td>rs1165278 4</td>
<td>0.000211</td>
<td>Whole blood</td>
</tr>
<tr>
<td>ZNF555</td>
<td>ENSG0000018630 0</td>
<td>rs887500 1</td>
<td>0.000364</td>
<td>Ovary</td>
</tr>
<tr>
<td>RPS26</td>
<td>ENSG0000019772 8</td>
<td>rs3741499 8</td>
<td>0.000378</td>
<td>Ovary</td>
</tr>
<tr>
<td>PTGER2</td>
<td>ENSG0000012538 4</td>
<td>rs781367 4</td>
<td>0.000403</td>
<td>Whole blood</td>
</tr>
<tr>
<td>RPS26</td>
<td>ENSG0000019772 8</td>
<td>rs3741499 8</td>
<td>0.000428</td>
<td>Whole blood</td>
</tr>
<tr>
<td>MAPKAP5-AS 1</td>
<td>ENSG0000023460 8</td>
<td>rs6489819 8</td>
<td>0.000465</td>
<td>Whole blood</td>
</tr>
<tr>
<td>UBE2Q2L</td>
<td>ENSG0000025951 1</td>
<td>rs2860012 1</td>
<td>0.000507</td>
<td>Ovary</td>
</tr>
<tr>
<td>ZNF556</td>
<td>ENSG0000018630 0</td>
<td>rs887500 1</td>
<td>0.000532</td>
<td>Ovary</td>
</tr>
<tr>
<td>ARMH2</td>
<td>ENSG0000026028 6</td>
<td>rs9393592 6</td>
<td>0.000589</td>
<td>Whole blood</td>
</tr>
<tr>
<td>RFWD3</td>
<td>ENSG0000016841 1</td>
<td>rs4072222 1</td>
<td>0.000731</td>
<td>Whole blood</td>
</tr>
<tr>
<td>MAST3</td>
<td>ENSG0000009930 8</td>
<td>rs1166628 8</td>
<td>0.000758</td>
<td>Whole blood</td>
</tr>
<tr>
<td>ICA1</td>
<td>ENSG0000000314 7</td>
<td>rs1095210 7</td>
<td>0.000763</td>
<td>Whole blood</td>
</tr>
<tr>
<td>RNF112</td>
<td>ENSG0000012848 2</td>
<td>rs1165278 2</td>
<td>0.000811</td>
<td>Whole blood</td>
</tr>
<tr>
<td>SERPINB10</td>
<td>ENSG0000024255 0</td>
<td>rs2162352 0</td>
<td>0.000971</td>
<td>Whole blood</td>
</tr>
<tr>
<td>HMGN1</td>
<td>ENSG0000020558 1</td>
<td>rs7279465 1</td>
<td>0.00113</td>
<td>Whole blood</td>
</tr>
<tr>
<td>ARHGAP1</td>
<td>ENSG0000017522 0</td>
<td>rs6485690 0</td>
<td>0.00122</td>
<td>Ovary</td>
</tr>
<tr>
<td>SLC41A1</td>
<td>ENSG0000013306 5</td>
<td>rs6681714 5</td>
<td>0.00126</td>
<td>Whole blood</td>
</tr>
</tbody>
</table>
Table 2 Common gene identified by TWAS and mRNA expression profiling

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>TWAS.Z</th>
<th>TWAS.P</th>
<th>Log2fold change</th>
<th>P.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYBRD1</td>
<td>-2.29174</td>
<td>0.02192</td>
<td>3.3603695</td>
<td>0.00470417</td>
</tr>
<tr>
<td>RAB31</td>
<td>-2.2727</td>
<td>0.023</td>
<td>1.6377338</td>
<td>0.00929241</td>
</tr>
<tr>
<td>NAGA</td>
<td>2.0323</td>
<td>0.04213</td>
<td>1.8187126</td>
<td>0.0084649</td>
</tr>
<tr>
<td>CAMK1D</td>
<td>3.01</td>
<td>0.00262</td>
<td>2.0792639</td>
<td>0.00507069</td>
</tr>
<tr>
<td>NUCKS1</td>
<td>3</td>
<td>0.0027</td>
<td>1.505372</td>
<td>0.00697271</td>
</tr>
<tr>
<td>PLCB2</td>
<td>2.98094</td>
<td>0.002874</td>
<td>2.3351678</td>
<td>0.00599101</td>
</tr>
<tr>
<td>COLEC12</td>
<td>2.7632</td>
<td>0.005724</td>
<td>1.4609694</td>
<td>0.00460956</td>
</tr>
<tr>
<td>TTC27</td>
<td>2.327847</td>
<td>0.0199</td>
<td>-1.1913658</td>
<td>0.00690472</td>
</tr>
<tr>
<td>GTF2IRD2</td>
<td>-2.28205</td>
<td>0.022486</td>
<td>3.3187845</td>
<td>0.00244768</td>
</tr>
<tr>
<td>LIMD2</td>
<td>-2.25126</td>
<td>0.024369</td>
<td>2.8804695</td>
<td>0.00075794</td>
</tr>
<tr>
<td>MS4A6A</td>
<td>-2.173</td>
<td>0.02978</td>
<td>1.8833963</td>
<td>0.00947628</td>
</tr>
<tr>
<td>PDE8B</td>
<td>2.15237</td>
<td>0.03137</td>
<td>-2.0226354</td>
<td>0.00865061</td>
</tr>
<tr>
<td>LIME1</td>
<td>2.151515</td>
<td>0.03144</td>
<td>1.2175341</td>
<td>0.0096489</td>
</tr>
<tr>
<td>RNASE2</td>
<td>2.0625</td>
<td>0.03916</td>
<td>2.7763519</td>
<td>0.00248818</td>
</tr>
<tr>
<td>LOC100505716</td>
<td>2.03598</td>
<td>0.0418</td>
<td>-2.1059969</td>
<td>0.00336991</td>
</tr>
<tr>
<td>KCNJ15</td>
<td>1.99801</td>
<td>0.04572</td>
<td>2.3330272</td>
<td>0.00287839</td>
</tr>
<tr>
<td>TGFBI</td>
<td>-1.97181</td>
<td>0.04863</td>
<td>3.2674614</td>
<td>0.00011652</td>
</tr>
<tr>
<td>CARMIL1</td>
<td>-1.9693</td>
<td>0.048924</td>
<td>-2.5048194</td>
<td>0.00130639</td>
</tr>
</tbody>
</table>
Figure 1 Manhattan plot of TWAS-identified top20 genes associated with PCOS

The x-axis represents the physical location of each gene, while the y-axis represents the Z score of association between that gene and PCOS.
Figure 2 Venn diagram of gene identified by TWAS and mRNA expression profiling
Figure 3 GO enrichment analysis of the genes identified by TWAS

GO:0034660: ncRNA metabolic process
GO:0009303: rRNA transcription
GO:0010970: transport along microtubule
GO:1901615: organic hydroxy compound metabolic process
GO:0007017: microtubule-based process
GO:0036297: interstrand cross-link repair
hsa00230: Purine metabolism
GO:0097352: autophagosome maturation
R-HSA:96945316: SARS-CoV-2 infection
GO:0016197: endosomal transport
GO:0071218: cellular response to misfolded protein
GO:0044283: small molecule biosynthetic process
GO:0126035: regulation of plasma membrane bounded cell projection organization
R-HSA:71387: Metabolism of carbohydrates
GO:0007764: small GTPase mediated signal transduction
GO:0006886: intracellular protein transport
R-HSA:5689877: josophin domain DUBs
GO:0195648: regulation of pinocytosis
R-HSA:1294445: Metabolism of porphyrins
WP4856: Intracellular trafficking proteins involved in CMT neuropathy
Figure 4 Protein-protein interaction network of genes identified by TWAS