


# Transcriptome-Wide Association Study Identifies Susceptibility Loci and Genes for Age at Natural Menopause

Reproductive Sciences  
2019, Vol. 26(4) 496-502  
© The Author(s) 2018  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/1933719118776788  
journals.sagepub.com/home/rsx  


Jiajun Shi, PhD<sup>1</sup>, Lang Wu, PhD<sup>1</sup>, Bingshan Li, PhD<sup>2,3</sup>, Yingchang Lu, MD, PhD<sup>1</sup>, Xingyi Guo, PhD<sup>1</sup>, Qiuyin Cai, MD, PhD<sup>1</sup>, Jirong Long, PhD<sup>1</sup>, Wanqing Wen, MD<sup>1</sup>, Wei Zheng, MD, PhD<sup>1</sup>, and Xiao-Ou Shu, MD, PhD<sup>1</sup> 

## Abstract

**Objective:** To identify novel susceptibility genes for age at natural menopause (ANM). **Methods:** Using transcription data generated in tissues from normal hypothalami ( $n = 73$ ) and ovaries ( $n = 68$ ) and high-density genotyping data provided by the Genotype-Tissue Expression (GTEx) database, we built 16 164 genetic models to predict gene expression across the transcriptome in these tissues. We used these models and summary statistics data from genome-wide association studies (GWAS) of ANM generated in 69 360 women of European ancestry to identify genes with their predicted expression related to ANM. **Results:** We found the predicted expression of 34 genes to be significantly associated with ANM at a Bonferroni-corrected threshold of  $P < 3.09 \times 10^{-6}$ . These include 4 genes located more than 1 Mb away from any previously GWAS-identified ANM-associated variants, 24 genes that reside in known GWAS-identified loci but have not been previously implicated, and 6 genes previously implicated as ANM-associated genes. **Conclusion:** Results from this transcriptome-wide association study, which integrated Expression quantitative trait loci (eQTL) data with summary statistics of GWAS of ANM, improves our understanding of the genetics and biology of female reproductive aging.

## Keywords

transcriptome-wide association studies, expression quantitative trait loci, genome-wide association studies, age at natural menopause

## Introduction

Age at natural menopause (ANM), occurring at an average age of 50 to 52 years among women of European ancestry, marks the end of a woman's normal reproductive life. Premature menopause and late age at menopause have both been linked to increased or reduced risk of various diseases such as cardiovascular diseases and breast cancer.<sup>1-6</sup> Recent genome-wide association studies (GWAS) among women of European ancestry have identified 54 independent single-nucleotide polymorphisms (SNPs) at 44 loci being associated with ANM.<sup>1,7-9</sup> However, the ANM-associated variants explain <6% of variance in ANM,<sup>1</sup> suggesting there are many additional genetic association signals for ANM yet to be identified.

A substantial proportion of the ANM-associated variants are located in non-protein-coding or intergenic regions.<sup>1</sup> It has been hypothesized that most of the GWAS-identified associations may be driven by the regulatory functions of specific identified variants or their tagged variants on the expression levels of genes that are involved in the physiology of complex traits or in the etiology of diseases.<sup>10-12</sup> For ANM, GWAS-identified associations at multiple loci were suggested to be

due to the effect of the variants in these loci on regulating the expression of approximately 25 genes.<sup>1,9</sup> For example, the index SNP rs365132, a synonymous variant in the *UIMC1* gene, regulates expression of both the coding gene itself and the downstream genes *ZNF346* (71 kb) and *FGFR4* (135 kb).<sup>1</sup> Yet, for the majority of the GWAS-identified ANM loci, the genes responsible for the associations remain undiscovered.

Recently, a transcriptome-wide association study (TWAS) approach has been developed to systematically investigate the association of genetically predictable gene expression with

<sup>1</sup> Department of Medicine, Vanderbilt Epidemiology Center and Division of Epidemiology, Vanderbilt University School of Medicine, Nashville, TN, USA

<sup>2</sup> Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN, USA

<sup>3</sup> Vanderbilt Genetics Institute, Nashville, TN, USA

## Corresponding Author:

Xiao-Ou Shu, Department of Medicine, Vanderbilt Epidemiology Center and Division of Epidemiology, Vanderbilt University School of Medicine, 2525 West End Avenue, Suite 600, IMPH, Nashville, TN 37203, USA.  
Email: xiao-ou.shu@vanderbilt.edu

complex traits. This approach has been shown to have the potential to identify the genes responsible for GWAS-identified associations for complex traits and diseases.<sup>13-17</sup> Instead of testing millions of SNPs in GWAS, TWAS evaluates the association of predicted expression for thousands of genes, greatly reducing the burden of multiple comparisons in statistical inference. A recent TWAS has successfully identified more than 200 candidate genes for age at menarche.<sup>18</sup> In the present study, we report results from the first large TWAS of ANM, involving 69 360 women of European ancestry.

## Materials and Methods

### Building of Gene Expression Prediction Models

We used transcriptome and high-density genotyping data from the Genotype-Tissue Expression (GTEx) database to establish prediction models for genes expressed in normal hypothalamus and ovary tissues. These tissues were chosen because they are particularly relevant to puberty timing and ovary aging.<sup>1</sup> Details of the GTEx have been described elsewhere.<sup>19</sup> Genomic DNA samples obtained from study participants included in the GTEx were genotyped using Illumina OMNI 5 M or 2.5 M SNP (Illumina, Inc., San Diego, CA, USA) Array, and RNA samples from different tissue sites were sequenced to generate transcriptome profiling data. We used genotyping and transcriptome data from women of European descent to build hypothalamus tissue ( $n = 73$ ) and ovary tissue ( $n = 68$ ) gene expression prediction models. Genotype data were processed according to the GTEx protocol (<http://www.gtexportal.org/home/documentationPage>). The SNPs with a call rate  $<98\%$ , with differential missingness between the 2 array experiments (5 M/2.5 M Arrays), with Hardy-Weinberg equilibrium  $P$  value  $<10^{-6}$  among participants of European ancestry, or showing batch effects, were excluded. One Klinefelter individual, 3 related individuals, and a chromosome 17 trisomy individual were also excluded. The genotype data were imputed to the Haplotype Reference Consortium reference panel<sup>20</sup> using Minimac3 for imputation and SHAPEIT for prephasing.<sup>21,22</sup> The SNPs with high imputation quality (RSQR  $\geq 0.8$ ), Minor allele frequency (MAF)  $\geq 0.05$ , included in the HapMap Phase 2 version, were used to build expression prediction models. For gene expression data, we used Reads Per Kilo-base per Million (RPKM) units from RNA-SeQC.<sup>23</sup> Genes with a median expression level of  $<0.1$  RPKM across samples were removed, and the RPKM values of each gene were log2 transformed. We performed quantile normalization to bring the expression profile of each sample to the same scale and performed inverse quantile normalization for each gene to map each set of expression values to a standard normal. We adjusted for the top 3 principal components (PCs) derived from genotype data and the top 15 probabilistic estimation of expression residual (PEER) factors to correct for batch effects and experimental confounders in model building.<sup>24</sup> Sex was adjusted during the gene expression process.

We built an expression prediction model for each gene with the elastic net method using the glmnet R package, with  $\alpha = .5$ ,

as recommended by Gamazon et al.<sup>13</sup> The genetically regulated expression for each gene was estimated by including variants within the 2 Mb flanking region of each gene. Expression prediction models were built for protein-coding genes, long non-coding RNA genes (lncRNAs), microRNAs (miRNAs), processed transcripts, immunoglobulin genes, and T cell receptor genes, according to categories described in the Gencode V19 annotation file (<http://www.encodegenes.org/releases/19.html>). Pseudogenes were not included in the present study because of potential concerns of inaccurate calling.<sup>25</sup> The 10-fold cross-validation strategy was used to validate the models internally. The prediction  $R^2$  values (ie, the square of the correlation between predicted and observed expression) were generated to estimate the prediction performance of each of the gene prediction models established.

### Summary Statistics of ANM GWAS

The summary statistics results of the GWAS of ANM were downloaded from the Reproductive Genetics Consortium (ReproGen) website ([http://www.reprogen.org/data\\_download.html](http://www.reprogen.org/data_download.html)) in December 2016. This GWAS comprised a maximum total sample size of 69 360 women of European descent, and the detailed information was described elsewhere.<sup>1</sup> Briefly, the GWAS comprised 33 individual studies using self-reported ANM. In each study, ANM associations were assessed using all autosomal SNPs imputed to reference panels of HapMap Phase 2 or 1000 Genomes Projects and under an additive model adjusted for top PCs and study-specific covariates. After standard quality control protocols, the study-specific results were then combined using an inverse variance-weighted meta-analysis. Only SNPs of no ambiguous strand (not A/T or C/G) and with MAF  $>0.01$  were used for the present study.

### Association Analyses of Predicted Gene Expression With ANM

We selected genes with a model prediction  $R^2$  of  $\geq .01$  in either ovary or hypothalamus tissues for association with ANM. Overall, a total of 16 164 models met the criteria and were evaluated for their expression-trait associations.

To identify ANM-associated genes, the MetaXcan method<sup>26</sup> was used for the association analyses. Briefly, the formula:

$$Z_g \approx \sum_{l \in \text{Model}_g} w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{\text{se}(\hat{\beta}_l)},$$

was used to estimate the Z-score of the association between predicted expression and ANM. Here,  $w_{lg}$  is the weight of SNP  $l$  for predicting the expression of gene  $g$ ,  $\hat{\beta}_l$  and  $\text{se}(\hat{\beta}_l)$  are the GWAS association regression coefficient and its standard error for SNP  $l$ ,  $\hat{\sigma}_l$  and  $\hat{\sigma}_g$  are the estimated variance of SNP  $l$  and the predicted expression of gene  $g$ , respectively. Therefore, the weights for predicting gene expression, GWAS summary statistics results, and correlations between

**Table 1.** Novel Expression Trait Associations for Genes Located in Genomic Loci not Previously Reported for Age at Natural Menopause.

Region	Gene	Model	Type	Z Score	P Value <sup>a</sup>	R <sup>2b</sup>
1q42.12	<i>EPHX1</i>	Hypothalamus	Protein	−4.80	$1.61 \times 10^{-6}$	0.02
5p15.31	<i>CTD-2044J15.2</i>	Ovary	LncRNA	−5.05	$4.53 \times 10^{-7}$	0.01
11q14.1	<i>RP11-452H21.1</i>	Hypothalamus	LncRNA	−4.91	$9.27 \times 10^{-7}$	0.04
19p12	<i>CTB-175P5.4</i>	Hypothalamus	LncRNA	4.69	$2.69 \times 10^{-6}$	0.33

Abbreviations: Protein, protein coding genes; LncRNA, long noncoding RNA genes.

<sup>a</sup> P value derived from association analyses, with  $P \leq 3.09 \times 10^{-6}$  considered statistically significant based on Bonferroni correction of 16 164 tests (0.05/16 164).

<sup>b</sup> R<sup>2</sup>: Prediction performance derived using GTEx data.

model-predicting SNPs are the input variables for the MetaXcan analyses. For this study, we estimated correlations between SNPs included in the prediction models using 1000 Genomes Project phase 3 data focusing on European population. We used a Bonferroni-corrected *P* threshold of  $3.09 \times 10^{-6}$  (0.05/16 164) to determine a statistically significant association for the primary analyses.

### Functional Enrichment Analysis Using INGENUITY Pathway Analysis

We performed functional enrichment analysis for our ANM TWAS-identified protein-coding genes. Canonical pathways, top associated diseases and biological functions, and top networks associated with genes of interest were estimated using the commercial Ingenuity Pathway Analysis (IPA) software.<sup>27</sup> These genes were analyzed for direct and indirect interactions in the IPA Knowledge Base. The gene enrichment significance for a pathway or a molecular function was assessed by a right-tailed Fisher exact test.<sup>9</sup> The significance of network association was expressed as a score, the negative log of the *P* value, denoting the likelihood of the input genes in a network being found together due to random chance.

## Results

### Gene Expression Prediction Models

The overall study design is shown in Supplementary Figure 1. Using GTEx data, we built hypothalamus tissue-based models for 10 807 genes, among which 7584 showed a prediction performance (*R*<sup>2</sup>) of at least .01 ( $\geq 10\%$  correlation between the predicted gene expression and measured gene expression; Supplementary Table 1). Of the 11 921 ovary tissue models we built, 8580 showed a prediction performance (*R*<sup>2</sup>) of at least .01 (Supplementary Table 2). Based on prior set criteria, we applied the 7584 hypothalamus tissue-based models and 8580 ovary tissue-based models for analysis of the association between predicted gene expression and ANM.

### Association Analyses of Predicted Gene Expression With ANM

Using the breast cancer GWAS data of 69 360 women of European ancestry and the 16 164 gene prediction models we built,

we evaluated the predicted gene expression levels for their associations with ANM. We identified 34 genes to be significantly associated with ANM at the Bonferroni-corrected threshold of  $P \leq 3.09 \times 10^{-6}$  (0.05/16 164; Tables 1-3). These include 4 genes located at 4 genomic loci that have not yet been reported to be associated with ANM (Table 1). Lower predicted expression of *EPHX1* (1q42.12), *CTD-2044J15.2* (5p15.31), and *RP11-452H21.1* (11q14.1) are associated with increasing ANM; conversely, higher predicted expression of *CTB-175P5.4* (19p12) is associated with increasing ANM (Table 1). The remaining 30 significantly associated genes are all located within 1 Mb from the GWAS-identified ANM-associated variants (Tables 2 and 3). However, the majority of these genes have not been reported as genes underlying these GWAS-identified SNPs and ANM associations (*n* = 24; Table 2). Only 6 protein-coding genes (*FGFR4*, *KRTCAP3*, *PRIM1*, *PRRC2A*, *RLBP1*, and *ZNF346*) have been previously reported as ANM-related genes through eQTL analyses of GWAS-identified ANM SNPs (Table 3).<sup>1,9</sup> Three other previously reported ANM-associated genes also showed nominal associations in the current analyses (under the hypothalamus models: *APT*X:  $P = 9.83 \times 10^{-5}$ ; *MSH6*:  $P = 1.87 \times 10^{-3}$ ; *RAD54L*:  $P = 4.17 \times 10^{-3}$ ; under the ovary models: *APT*X:  $P = .046$ ; *MSH6*:  $P = .031$ ). For these 9 genes, the association identified in our study showed the same direction with that previously implicated based on eQTL data and the GWAS-identified variant association (data not shown). Although the literature has suggested some other candidate ANM-associated genes, which were not identified in this study, most of them were based on eQTL analyses using expression data from tissues such as blood, skin, adipose, and prefrontal cortex.<sup>1,9</sup> In this study, we focused on the tissues most relevant to ANM (ovary and hypothalamus).

To understand the relationship between our identified associations and known GWAS-identified variants for ANM (Tables 2 and 3), we used a method proposed by Yang et al<sup>28</sup> to perform conditional analyses adjusting for the index SNPs in each of the known ANM-associated regions. We found that after adjusting for multiple comparisons (0.05/30), the associations of 5 genes with ANM remained statistically significant at  $P < 1.67 \times 10^{-3}$  (Tables 2 and 3), suggesting that these genes were associated with ANM at least partially independent of the index SNPs. In fact, the association of *TRAPPC2L* with ANM remained significant at  $P \leq 3.09 \times 10^{-6}$  (Table 2), displaying

**Table 2.** Novel Expression Trait Associations for Genes in Genomic Loci Previously Reported for Age at Natural Menopause.

Region	Gene	Model	Type	Z Score	P Value <sup>a</sup>	R <sup>2b</sup>	Index SNP(s) <sup>c</sup>	Distance to the Index SNP (kb)	P Value after Adjusting for Index SNPs <sup>d</sup>
2p23.3	<i>ATRAID</i>	Hypothalamus	Protein	-5.47	$4.40 \times 10^{-8}$	0.04	rs2303369	275	.001
2q31.1	<i>AC068039.4</i>	Ovary	LncRNA	-6.61	$3.94 \times 10^{-11}$	0.26	rs10183486	639	$2.02 \times 10^{-5}$
		Hypothalamus	LncRNA	-4.90	$9.62 \times 10^{-7}$	0.19	rs10183486	639	$9.20 \times 10^{-6}$
5q35.2	<i>HIGD2A</i>	Hypothalamus	Protein	5.34	$9.44 \times 10^{-8}$	0.03	rs890835	139	$2.30 \times 10^{-4}$
	<i>GPRIN1</i>	Hypothalamus	Protein	-7.58	$3.52 \times 10^{-14}$	0.14	rs890835	67	.01
5q35.3	<i>RP11-1334A24.6</i>	Hypothalamus	LncRNA	7.42	$1.17 \times 10^{-13}$	0.01	rs365132	543	$3.68 \times 10^{-4}$
6p21.33	<i>TCF19</i>	Hypothalamus	Protein	6.69	$2.16 \times 10^{-11}$	0.07	rs1046089	468	.7
	<i>BAG6</i>	Hypothalamus	Protein	6.14	$8.25 \times 10^{-10}$	0.01	rs1046089	4	.23
	<i>CSNK2B</i>	Hypothalamus	Protein	-4.93	$8.39 \times 10^{-7}$	0.12	rs1046089	30	.06
6p21.33	<i>NOTCH4</i>	Hypothalamus	Protein	4.77	$1.88 \times 10^{-6}$	0.01	rs494620	324	.78
9p21.1	<i>AL162590.1</i>	Hypothalamus	miRNA	5.26	$1.42 \times 10^{-7}$	0.28	rs4879656	29	.18
12q13.2	<i>SUOX</i>	Hypothalamus	Protein	-4.76	$1.97 \times 10^{-6}$	0.08	rs2277339	746	.03
	<i>RP11-603J24.17</i>	Hypothalamus	LncRNA	-5.06	$4.09 \times 10^{-7}$	0.04	rs2277339	638	.04
	<i>ZBTB39</i>	Hypothalamus	Protein	-6.99	$2.70 \times 10^{-12}$	0.01	rs2277339	247	.07
12q24.31	<i>OGFOD2</i>	Hypothalamus	Protein	-5.49	$4.02 \times 10^{-8}$	0.17	rs1727326	135	.07
15q15.1	<i>OIP5-AS1</i>	Ovary	Transcript	-5.27	$1.39 \times 10^{-7}$	0.15	rs9796	305	.16
	<i>NUSAPI</i>	Hypothalamus	Protein	-5.31	$1.13 \times 10^{-7}$	0.42	rs9796	353	$2.37 \times 10^{-5}$
16q24.3	<i>TRAPPC2L</i>	Ovary	Protein	5.25	$1.54 \times 10^{-7}$	0.05	rs4843747	899	$3.09 \times 10^{-9}$
17q21.2	<i>RAB5C</i>	Hypothalamus	Protein	5.74	$9.22 \times 10^{-9}$	0.04	rs1799949	938	.006
19q13.42	<i>SSC5D</i>	Hypothalamus	Protein	4.89	$1.00 \times 10^{-6}$	0.04	rs11668344	166	.34
20p12.3	<i>TMEM230</i>	Hypothalamus	Protein	-5.22	$1.82 \times 10^{-7}$	0.23	rs236114	842	.01
22q13.1	<i>FAM227A</i>	Hypothalamus	Protein	-6.93	$4.07 \times 10^{-12}$	0.12	rs763121	94	.04
	<i>CBY1</i>	Ovary	Protein	5.07	$3.92 \times 10^{-7}$	0.38	rs763121	173	.14
	<i>RP3-508115.10</i>	Ovary	LncRNA	6.21	$5.20 \times 10^{-10}$	0.17	rs763121	183	.03
	<i>RP3-508115.9</i>	Ovary	LncRNA	6.62	$3.53 \times 10^{-11}$	0.24	rs763121	184	.16

Abbreviations: protein, protein coding genes; lncRNA, long noncoding RNA genes; transcript, processed transcript; miRNA, microRNA genes; SNP, single-nucleotide polymorphism.

<sup>a</sup>P value derived from association analyses; associations with  $P \leq 3.09 \times 10^{-6}$  considered statistically significant based on Bonferroni correction of 16 164 tests (0.05/16 164).

<sup>b</sup>R<sup>2</sup>: Prediction performance derived using GTEx data.

<sup>c</sup>Index SNPs identified in previous genome-wide association studies (GWAS) or fine-mapping studies. When multiple risk variants were identified in the same region, the SNP closest to the gene is presented.

<sup>d</sup>Use of COJO method<sup>28</sup>; all index SNPs in the corresponding region were adjusted for the conditional analyses.

**Table 3.** Significant Expression Trait Associations for Genes Previously Implicated for Age at Natural Menopause.

Region	Gene	Model	Type	Z Score	P Value <sup>a</sup>	R <sup>2b</sup>	Index SNP(s) <sup>c</sup>	Distance to the Index SNP (kb)	P Value after Adjusting for Index SNPs <sup>d</sup>	Association Direction Reported Previously	Reference
2p23.3	<i>KRTCAP3</i>	Hypothalamus	Protein	-7.14	$9.33 \times 10^{-13}$	0.06	rs2303369	46	0.003	NA	8
5q35.2	<i>ZNF346</i>	Hypothalamus	Protein	-6.17	$6.88 \times 10^{-10}$	0.06	rs365132	71	0.94	-	8,9
	<i>FGFR4</i>	Hypothalamus	Protein	5.03	$4.94 \times 10^{-7}$	0.12	rs365132	135	0.48	-	8
6p21.33	<i>PRRC2A</i>	Hypothalamus	Protein	6.17	$6.67 \times 10^{-10}$	0.04	rs1046089 <sup>e</sup>	Within the gene	0.07	NA	8
12q13.3	<i>PRIMI</i>	Ovary	Protein	4.86	$1.19 \times 10^{-6}$	0.04	rs2277339 <sup>e</sup>	Within the gene	0.68	NA	8,9
15q26.1	<i>RLBPI</i>	Hypothalamus	Protein	-4.93	$8.44 \times 10^{-7}$	0.02	rs2307449	99	0.39	NA	8

Abbreviations: protein, protein coding genes; SNP, single-nucleotide polymorphism; NA, Not Available.

<sup>a</sup>P value derived from association analyses, with  $P \leq 3.09 \times 10^{-6}$  considered statistically significant based on Bonferroni correction of 16 164 tests (0.05/16 164).

<sup>b</sup>R<sup>2</sup>: Prediction performance derived using GTEx data.

<sup>c</sup>Index SNPs identified in previous genome-wide association studies (GWAS) or fine-mapping studies. When multiple risk variants were identified in the same region, the SNP closest to the gene is presented.

<sup>d</sup>Use of COJO method<sup>28</sup>; all index SNPs in the corresponding region were adjusted for the conditional analyses.

<sup>e</sup>Index SNPs rs1046089 and rs2277339 are predicted damaging missense variants in the genes *PRRC2A* and *PRIMI*, respectively.

strong evidence of association independently from the index SNPs at 16q24.3.

### Pathway Analyses

The IPA suggested that 10 of the 24 identified ANM protein-coding genes were within 1 major network related to embryonic and organismal development (Supplementary Table 3). These genes are significantly enriched in several canonical pathways such as clathrin-mediated endocytosis signaling, interleukin 2 (IL-2) signaling, and epidermal growth factor (EGF) signaling (Supplementary Table 3).

### Discussion

In this TWAS, we systematically evaluated genetically predicted gene expression in hypothalamus and ovary tissues for their association with ANM and identified 34 genes showing a significant association with ANM at the Bonferroni-corrected threshold. Among them, 28 genes, including 4 genes located at 4 novel genomic loci (*EPHX1* at1q42.12, *CTD-2044J15.2* at 5p15.31, *RP11-452H21.1* at 11q14.1, and *CTB-175P5.4* at 19p12) and 24 genes within the previous GWAS-identified ANM loci, have not been reported to be associated with ANM.

Several of the protein-coding genes identified in our study have previously been implicated in female reproductive aging. The DNA polymerase *PRIM1* synthesizes small RNA primers for the Okazaki fragments made during discontinuous DNA replication,<sup>29</sup> and its coding gene is one of the DNA damaging genes known to be associated with ovary aging.<sup>1</sup> Epoxide hydrolase 1, encoded by the *EPHX1* gene, plays a role in estrogen production in the human ovary.<sup>30</sup> Methylation of the *EPHX1* promoter is associated with polycystic ovary syndrome.<sup>31</sup> We found a significant association between decreased *EPHX1* expression in the hypothalamus and an older ANM, further supporting possible links between *EPHX1* and ovary development and reproductive aging. In this study, we also identified 8 lncRNAs to be associated with ANM. Although lncRNAs have been shown to play important roles in the regulation of gene expression, cell biology, and cancer development, including carcinogenesis of breast cancer,<sup>32,33</sup> the molecular and biological functions of lncRNAs in ovary aging have not been previously reported and our interesting findings warrant further investigation.

To our knowledge, this is the first large study using a TWAS design to identify candidate susceptibility genes of ANM. We used data from the largest available GWAS of ANM in our analyses, providing high statistical power for the association analysis. Unlike a typical GWAS study, we have been able to provide information on direction of the association for the identified genes by evaluating the associations of predicted gene expression, which could facilitate future functional investigations. On the other hand, several potential limitations need to be considered to appropriately interpret our findings. First, although we have used the conservative Bonferroni-corrected threshold to minimize type 1 error, we acknowledge that

false-positive associations may still exist in our TWAS. Our approach could not completely exclude the possibility that the identified genes are associated with ANM merely through a linkage disequilibrium between SNPs predicting the expression of corresponding genes and a causal SNP of ANM acting through an alternative mechanism (so-called LD contamination). Co-regulation (ie, a variant regulating the expression of multiple genes) may also have resulted in false positives in the study. We identified several genes colocalized at the same locus (eg, 12q13.2 locus) showing a same direction of effect but were not able to pinpoint which of them is the truly causal gene. Future functional studies evaluating functional significance of our identified genes in menopausal timing are needed. Second, the sample sizes for building gene expression prediction models of normal hypothalamus and ovary tissues were relatively small, which could affect the precision of parameter estimates of the built models. Prediction models built with a larger sample size in future efforts would help identify additional candidate genes associated with ANM. Furthermore, although our findings are based on the largest available GWAS to date, future efforts to use data from additional resources, for example, the UK Biobank,<sup>34</sup> would be necessary to uncover additional ANM genes and replicate our identified associated genes.

Primary ovarian insufficiency (POI) or premature ovarian failure (POF), defined by menopause before the age of 40, can have significant physical and psychological impacts on affected women. Previous studies have reported multiple genes associated with POI/POF.<sup>1,35,36</sup> However, none of these genes were found to be related to ANM in our study, probably due to low statistical power, as POI/POF only occurs in approximately 1% of all women.<sup>37</sup> Lack of information on POI/POF prevented us from investigating this phenotype.

In summary, we performed the first TWAS of ANM, resulting in the identification of multiple genes associated with ANM. These genes are enriched in the embryo and organism development network and are likely to be involved in ovarian aging. Results from our study improve the understanding of the genetics and biology of female reproductive aging.

### Authors' Note

Jiajun Shi, PhD, and Lang Wu, PhD, contributed equally to this study.

X.O.S. and W.Z. conceived the study. L.W. and J.S. contributed to the study design and performed the statistical analyses. J.S. and L.W. drafted the manuscript with significant contributions from X.O.S. and W. Z. Y. L. contributed to the model building. X.G. contributed to the pathway analyses. All authors provided suggestions during the data analyses, participated in data interpretation, and critically reviewed and approved the final manuscript.

The Genotype-Tissue Expression (GTEx) Project data were obtained from the GTEx Portal (<https://www.gtexportal.org/home/datasets>). Institution of Research: Vanderbilt University School of Medicine.

### Acknowledgment

The authors wish to thank the Reproductive Genetics Consortium for making the GWAS summary statistics data publicly available (<http://>

www.reprogen.org/data\_download.html). The GTEx dataset version phs000424.v6.p1 was used in the study under the dbGaP-approved protocol #22412-8. The authors acknowledge and thank the Genotype-Tissue Expression (GTEx) consortium for making data publicly available via dbGaP ([www.ncbi.nlm.nih.gov/gap](http://www.ncbi.nlm.nih.gov/gap)). We also thank Ms. Nan Kennedy, Division of Epidemiology, Vanderbilt University Medical Center, for her assistance in preparing the article.

### Declaration of Conflicting Interests

The author(s) declared no potential conflict of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The study was supported, in part, by grants from the National Institutes of Health (R01CA090899, R01CA064277, R01CA118229, R01CA148667) and Ingram professorship funds and Allen Foundation funds. Lang Wu was supported by the Vanderbilt Molecular and Genetic Epidemiology of Cancer (MAGEC) training program funded by the US National Cancer Institute grant R25 CA160056 (PI: X.-O. Shu).

### ORCID iD

Xiao-Ou Shu, MD, PhD  <https://orcid.org/0000-0002-0711-8314>

### Supplemental Material

Supplementary material for this article is available online.

### References

- Day FR, Ruth KS, Thompson DJ, et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat Genet.* 2015;47(11):1294-1303.
- Hinds L, Price J. Menopause, hormone replacement and gynaecological cancers. *Menopause Int.* 2010;16(2):89-93.
- Parker SE, Troisi R, Wise LA, et al. Menarche, menopause, years of menstruation, and the incidence of osteoporosis: the influence of prenatal exposure to diethylstilbestrol. *J Clin Endocrinol Metab.* 2014;99(2):594-601.
- Qiu C, Chen H, Wen J, et al. Associations between age at menarche and menopause with cardiovascular disease, diabetes, and osteoporosis in Chinese women. *J Clin Endocrinol Metab.* 2013;98(4):1612-1621.
- Velie EM, Nechuta S, Osuch JR. Lifetime reproductive and anthropometric risk factors for breast cancer in postmenopausal women. *Breast Dis.* 2005;24:17-35.
- Vogel VG. Epidemiology, genetics, and risk evaluation of postmenopausal women at risk of breast cancer. *Menopause.* 2008; 15(4 suppl):782-789.
- He C, Kraft P, Chen C, et al. Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. *Nat Genet.* 2009;41(6):724-728.
- Stolk L, Zhai G, van Meurs JBJ, et al. Loci at chromosomes 13, 19 and 20 influence age at natural menopause. *Nat Genet.* 2009; 41(6):645-647.
- Stolk L, Perry JRB, Chasman DI, et al. Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways. *Nat Genet.* 2012;44(3):260-268.
- Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 2010;6(4): e1000888.
- Nica AC, Montgomery SB, Dimas AS, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* 2010;6(4): e1000895.
- Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet.* 2015;16(4):197-212.
- Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet.* 2015;47(9):1091-1098.
- Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet.* 2016; 48(3):245-252.
- Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet.* 2016;48(5):481-487.
- Pavlidis JMW, Zhu Z, Gratten J, McRae AF, Wray NR, Yang J. Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. *Genome Med.* 2016;8(1):84.
- Mancuso N, Shi H, Goddard P, Kichaev G, Gusev A, Pasaniuc B. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am J Hum Genet.* 2017;100(3):473-487.
- Day FR, Thompson DJ, Helgason H, et al. Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nat Genet.* 2017; 49(6):834-841.
- GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science.* 2015;348(6235):648-660.
- McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet.* 2016; 48(10):1279-1283.
- Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 2009;5(6):e1000529.
- Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. *Nat Methods.* 2011;9(2): 179-181.
- DeLuca DS, Levin JZ, Sivachenko A, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics Oxf Engl.* 2012;28(11):1530-1532.
- Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc.* 2012;7(3):500-507.
- Guo X, Lin M, Rockowitz S, Lachman HM, Zheng D. Characterization of human pseudogene-derived non-coding RNAs for functional potential. *PLoS One.* 2014;9(4):e93972.

26. Barbeira A, Shah KP, Torres JM, et al. MetaXcan: summary statistics based gene-level association method infers accurate predixcan results. *bioRxiv*. 2016; <https://doi.org/10.1101/045260>
27. Krämer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinforma Oxf Engl*. 2014;30:523-530.
28. Yang J, Ferreira T, Morris AP, Genetic Investigation of ANthropometric Traits (GIANT) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet*. 2012;44(4):369-375, S1-S3.
29. Shiratori A, Okumura K, Nogami M, et al. Assignment of the 49-kDa (PRIM1) and 58-kDa (PRIM2A and PRIM2B) subunit genes of the human DNA primase to chromosome bands 1q44 and 6p11.1-p12. *Genomics*. 1995;28(2):350-353.
30. Hattori N, Fujiwara H, Maeda M, Fujii S, Ueda M. Epoxide hydrolase affects estrogen production in the human ovary. *Endocrinology*. 2000;141(9):3353-3365.
31. Sang Q, Li X, Wang H, et al. Quantitative methylation level of the EPHX1 promoter in peripheral blood DNA is associated with polycystic ovary syndrome. *PLoS One*. 2014;9(2):e88013.
32. Cerk S, Schwarzenbacher D, Adiprasito JB, et al. Current status of long non-coding RNAs in human breast cancer. *Int J Mol Sci*. 2016;17(9). pii: E1485.
33. Evans JR, Feng FY, Chinnaiyan AM. The bright side of dark matter: lncRNAs in cancer. *J Clin Invest*. 2016;126(8):2775-2782.
34. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779.
35. Chapman C, Cree L, Shelling AN. The genetics of premature ovarian failure: current perspectives. *Int J Womens Health*. 2015;7:799-810.
36. Rossetti R, Ferrari I, Bonomi M, Persani L. Genetics of primary ovarian insufficiency. *Clin Genet*. 2017;91(2):183-198.
37. Shelling AN. Premature ovarian failure. *Reprod Camb Engl*. 2010;140(5):633-641.